# A multiple-phenotype imputation method for genetic studies

# A multiple phenotype imputation method for genetic studies

Andrew Dahl[1,8], Valentina Iotchkova[2,3,8], Amelie Baud[3], Åsa Johansson[4], Ulf Gyllensten[4], Nicole Soranzo[2], Richard Mott[1], Andreas Kranis[5,6], Jonathan Marchini[7,1]

[1] The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

[2] Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

[3] The European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, UK

[4] Department of Immunology, Genetics, and Pathology, SciLifeLab Uppsala, Uppsala University, Uppsala, Sweden

[5] Aviagen Ltd, Newbridge, United Kingdom

[6] The Roslin Institute, University of Edinburgh, Midlothian, United Kingdom

[7] Department of Statistics, University of Oxford, Oxford, UK

[8] Joint first authors of this paper

**Correspondence:**

**Professor Jonathan Marchini**

Department of Statistics

University of Oxford

1 South Parks Road

Oxford OX1 3TG, UK

Tel: +44 (0)1865 271125

Fax: +44 (0)1865 281333

E-mail: marchini@stats.ox.ac.uk

**Abstract**

Genetic association studies have yielded a wealth of biologic discoveries. However, these have mostly analyzed one trait and one SNP at a time, thus failing to capture the underlying complexity of these datasets. Joint genotype-phenotype analyses of complex, high-dimensional datasets represent an important way to move beyond simple GWAS with great potential. The move to high-dimensional phenotypes will raise many new statistical problems. In this paper we address the central issue of missing phenotypes in studies with any level of relatedness between samples. We propose a multiple phenotype mixed model and use a computationally efficient variational Bayesian algorithm to fit the model. On a variety of simulated and real datasets from a range of organisms and trait types, we show that our method outperforms existing state-of-the-art methods from the statistics and machine learning literature and can boost signals of association.

## Introduction

Genome-wide association studies (GWAS) have successfully uncovered many associated loci. Such approaches typically analyze thousands of nominally unrelated individuals and search for correlations between genetic variants and a single trait of interest. However, a complete characterization of the etiology of most traits remains elusive. This may be because the GWAS approach is quite crude, in that much of the biology between sequence and phenotype remains unmeasured. Large scale phenotyping is starting to generate invaluable data that can be harnessed by geneticists [1].

This observation motivates the analysis of multiple phenotypes, traits and sub-phenotypes, a direction that is increasingly prominent in the literature of human, plant and animal genetics [2-5] . The advantages of analyzing multiple phenotypes related to, or underlying, a phenotype of interest include boosting power to detect novel associations[6], measuring heritable covariance between traits [7] and the potential to make causal inference between traits [8].

At the same time, harnessing genetic relatedness, even amongst nominally unrelated samples, to boost power in association studies is becoming increasingly prevalent. Mixed models, re-emerging from the linkage and animal genetics literature[9-11] , are now routinely used to search for associations in the presence of relatedness or population structure and to estimate the additive genetic component of heritability. However, until recently these analyses have mostly proceeded one trait at a time.

In this paper, we consider the analysis of multiple correlated phenotypes observed on correlated samples, which arises with related individuals, cryptic relatedness, population structure or polygenicity. Crucially, the vast majority of methods for multiple phenotypes rely on all samples having fully observed phenotypes[3,6]. However, as the number of phenotypes increases the chance that at least one observation is missing increases exponentially. Removal of all samples with a missing phenotype will reduce sample size, thus attenuating the

power of any statistical inference. For example, a range of real studies removed between 3%-31%[12-17] of samples. Other studies completely removed phenotypes with high levels of missing data, and imputed remaining missing data with off-the-shelf methods from mainstream statistics [18-20]. While re-phenotyping of samples is ideal, it is typically expensive or infeasible [21].

We propose a method to impute missing phenotypes in related samples, which will likely be a crucial first step for many downstream analyses. In this setting correlations will exist between phenotypes and between samples, and *both* are useful in predicting missing observations. We propose a Bayesian multiple phenotype mixed model and use a Variational Bayesian (VB) method to fit the model. We assume that the kinship between individuals in a study is known *a priori* from genetic data[22] or a pedigree. This information enables the model to decompose the correlation between traits into a genetic and a residual component. A notable feature of our method is that it can handle hundreds of traits. We call our method PHENIX.

We validate our approach with an extensive simulation study, representative of a variety of genetic studies of humans and other organisms. We compare our method to approaches that ignore either the correlations between samples or the correlations between traits, and to state-of-the-art missing data imputation techniques from mainstream statistics and machine learning. We also apply our method to five real datasets on a variety of traits from humans[2,23], yeast[24], rats[25], chickens [26] and wheat [27]. In all simulated and real datasets we show evidence that our method outperforms the competing methods in accuracy and is computationally efficient. We also apply the method to a rat GWAS of 140 phenotypes to illustrate how the method can be used to boost signals of association. Finally, we discuss the usefulness of this approach, the range of relevant datasets that the method could be applied to, and how the method might be developed further in the future.

## Results

Simulations

We simulated datasets with N=300 individuals and P=15 traits varying the level of relatedness between individuals and the heritability of the traits. A standard multiple phenotype mixed model (MPMM) was used to simulate phenotypes with an underlying genetic covariance, as well as added environmental, or residual, correlation. For the genetic covariance between traits we used a model with a range of positive and negative correlations between the traits. For the residual covariance we added randomly correlated noise to the phenotypes. We varied the heritability of the traits by adjusting the relative contributions of the genetic and residual covariance terms. We used two models for relatedness between samples: Model 1 used an empirical kinship matrix derived from the Northern Sweden Population Health Study (NSPHS) [23]; Model 2 simulated 75 independent families of 4 full siblings. Missing data was added completely at random at the 5% level. The true values of missing data were kept to measure performance. We averaged results over 100 datasets simulated under each scenario. More details are given in the **Online Methods**.

We fit our method (PHENIX) to each of the simulated datasets to infer point estimates of the missing phenotypes. We assessed performance by measuring the correlation between these imputed phenotypes and their true hidden values. The results are shown in **Figure 1** for both levels of relatedness. We compared our method to a range of other imputation methods from the statistical genetics, mainstream statistics and machine learning literatures (**Table 1, Online methods**). These methods model different aspects of the correlation structure in the data, in most cases ignoring genetic or phenotypic correlations; PHENIX models both aspects. Results using a mean squared error (MSE) metric and timing information are shown in **Supplementary Figure 1** and the **Supplementary Note,** respectively**.**

The overall pattern from **Figure 1** is that PHENIX outperforms all other methods over the full range of heritability. As heritability increases the difference

between PHENIX and the next best method increases. A number of other interesting patterns also emerge. Ignoring correlations between phenotypes (LMM – green line) is mostly a much worse assumption than ignoring correlations between samples (MVN – blue line), except at very high heritabilities and high levels of relatedness between samples (Model 2). In fact, ignoring correlations between samples does remarkably well, especially considering MVN is the fastest method in our comparisons. However, the performance of MVN suffered in some real datasets with high relatedness (**Figure 2**) so we do not recommend it for general use. TRCMA (pink line) and SOFTIMPUTE (cyan line) seem to perform roughly equally well, and better than MICE and kNN (brown and grey lines respectively). This is likely because the former two methods partially model sample relatedness, whereas the latter two methods only model phenotypic correlations. Most methods were fast enough to be practical, although we found TRCMA to be prohibitively slow in most settings (**Supplementary Note**).

Increasing levels of relatedness between samples increases the accuracy of PHENIX and LMM. Both of these methods explicitly take account of the relatedness between samples via the kinship matrix. For example, when the heritability of the traits is $h^2$=0.3, the imputation correlation of PHENIX is 0.63 and 0.67 on Model 1 (NSPHS) and Model 2 (sibs) respectively.

As heritability increases the performance of all the best performing methods decreases, but then increases slightly again as heritability approaches 1. This occurs because the overall correlations between traits are a mixture of genetic and environmental correlations. At intermediate heritability the genetic and environmental correlations tend to cancel each other out, attenuating the performance of methods that harness phenotypic correlations. To highlight this effect we carried out simulations in which genetic and environmental covariances are the inverses of each other. At intermediate values of heritability the performance of all methods suffers (**Supplementary Figure 2**).

When the number of samples is increased to N=1000 and phenotypes to P=50 the performance of PHENIX improves compared to the other methods, especially for Model 1 which uses an empirical kinship matrix derived from the NSPHS study (**Supplementary Figure 3**). As the genetic correlation between traits increases, the residual contribution becomes less important and thus the utility from partitioning the covariance is attenuated; this means the gap between PHENIX and other methods shrinks. Conversely, when the genetic correlation shrinks, PHENIX increasingly outperforms the others (**Supplementary Figure 4**). Increasing the missing data rate to 10% degrades performance for all methods, especially when there are few close relationships between samples (**Supplementary Figure 5**). We investigated the effects of non-random missingness (**Supplementary Figure 6**), unmodelled, shared environmental effects (**Supplementary Figure 7**) and non-normally distributed phenotypes (**Supplementary Figure 8**), which all act to reduce performance in general. However, PHENIX remains the best performing method in all scenarios.

A likely main use of PHENIX is to impute missing phenotypes ahead of association testing of phenotypes with genome-wide SNP data. This might proceed by testing phenotypes one at a time, or by using a multi phenotype association test. As such it is important to show that our approach leads to valid statistical tests. Using simulated phenotype data and real genotype data from the NSPHS cohort (described below) we find that association testing after imputation results in well calibrated p-values under the null (**Supplementary Figure 9**).

There is a large literature on multi-phenotype tests [3,6,28-30] and there seems wide consensus that these tests can lead to an increase in power over single phenotype tests in many realistic scenarios. We assessed whether imputing missing phenotypes can increase in power when testing a SNP for association. We find that imputation can lead to an increase in power when testing either one phenotype at a time, or when using a multi-phenotype test (**Supplementary Figures 10 and 11**)**.** Intuitively, one of the main reasons this occurs is that imputation increases the sample size used in the test.

Real data

To further illustrate the usefulness of PHENIX we imputed missing phenotypes in several real datasets. We applied the method to a range of different organisms to illustrate that our method will be useful in a wide variety of settings and across a diverse set of phenotypes used in real genetic studies. Animal and plant studies almost always use related samples, due to study design constraints, but in some cases, like Arabidopsis, unrelated samples with considerable population structure are used.

The datasets are hematological measurements in the UK Blood Services Common Control (UKBS) collection that was studied by the HaemGen consortium [2], glycans phenotypes in the NSPHS study[4,23], phenotypes related to six disease models and measures of risk factors for common diseases in outbred rats[25], phenotypes measuring growth of yeast under different conditions [24], phenotypes relevant to a genomic selection program in a multigenerational chicken pedigree[26] and traits related to growth and yield in an inter-cross population for winter-sown wheat[27]. **Table 2** details the properties of these datasets.

Each of these datasets has a different level of missing data. We created new datasets by increasing levels of missing data, keeping the true values to assess imputation performance. We applied the various imputation methods to these datasets and measured performance using the correlation between the imputed and true values. The results for each of the six datasets are presented in **Figure 2**, where imputation correlation (y-axis) is plotted against missing data percentage (x-axis). The true level of missing data is highlighted as a vertical, dashed black line.

As in the simulated datasets, PHENIX is the most accurate method across all six of the datasets, except at extreme levels of missingness. For realistic levels of missing data, near the actual levels in the datasets, PHENIX clearly outperforms the other methods in the yeast and chicken datasets, but the difference is smaller on the human, rat and wheat datasets. On all six datasets TRCMA, SOFTIMPUTE

and MVN perform almost the same. As with the simulated data, these 3 methods tend to outperform MICE, which in turn tends to outperform kNN.

The single trait LMM method is overall the worst performing method, however it does reasonably well on the yeast and chicken datasets, where the trait heritabilities and levels of sample relatedness are high and traits are relatively uncorrelated compared to the other datasets. Appropriately, these are the datasets where PHENIX substantially outperforms TRCMA, SOFTIMPUTE and MVN.

For the human NSPHS and wheat datasets we fit a standard Multiple Phenotype Mixed Model (MPMM), with an EM algorithm[31], only to those individuals with fully observed phenotypes, and used the estimated parameters to impute missing phenotypes in other individuals, following others[3]. MPMM will not run on the human UKNBS, yeast, chicken or rat datasets where the number of phenotypes and levels of missingness produce no samples with complete observations. When it is possible to apply this method we observed (**Figure 2 – purple lines**) that its performance drops off considerably. As the amount of missing data increases the number of samples with completely observed phenotypes will exponentially decrease, which will harm parameter estimation and subsequent imputation performance.

Application to Rat GWAS

To assess the utility of our method in the GWAS setting we re-analyzed the data from the Rat Genome Sequencing and Mapping Consortium. Specifically, we imputed all the missing phenotypes and covariates available in the deposited dataset. We then carried out GWAS for the 140 most biologically relevant phenotypes (those mapped in the original study[25]) at the 24,196 genomic locations at which HAPPY[32] descent probabilities had been calculated (see **Online Methods**). The amount of missing data in these 140 phenotypes varies from 1.5% to 87% (median=16.6%). We then compared these results to GWAS performed on the phenotypes without imputation.

In much the same way that information scores are used when carrying out downstream analyses such as GWAS on imputed genotypes[33], it is desirable to assess the accuracy of phenotype imputation. To achieve this, we added extra missing data, re-imputed the missing phenotypes and then calculated an imputation squared correlation ($r^2$) for each phenotype using the held out data (see **Online Methods**). This metric can be automatically calculated by the imputation functions in our `R` package, and experiments suggest that the measure is very accurately calibrated (**Supplementary Figure 12**). To choose a useful threshold for $r^2$, we used experience of filtering genotype imputation information scores, which typically filter at some value between 0.3-0.4. Ultimately, we used 82 phenotypes with $r^2 > 0.36$. The amount of missing data being imputed may also be a useful phenotype summary to consider when interpreting imputation results.

**Figure 3** compares the results of the imputed and un-imputed rat GWAS for all 140 phenotypes. To report results we applied a conservative p-value threshold of $-\log_{10}(p) > 10$. We only plot p-values for genomic locations that are maximal in a 6 Mb window (+/- 3 Mb). These are plotted against the maximum $-\log_{10}(p)$ in the same 6 Mb window in the complementary (imputed or un-imputed) GWAS. Grey points are those for which $r^2 < 0.36$. The cluster of grey points with imputed $-\log_{10}(p) < 10$ and un-imputed $-\log_{10}(p) > 10$ all correspond to phenotypes with very low $r^2$ and high levels of missing data demonstrating that filters on $r^2$ and missingness can identify when imputation results should be viewed with caution.

The figure highlights that there are circumstances where phenotype imputation has a good imputation $r^2$ and acts to increase the signal of association (red and blue points). A cluster of associations (red points) all correspond to three related platelet phenotypes (mean platelet volume (MPV), mean platelet count (MPC) and platelet distribution width (PDW)) over an extended region of chromosome 9 between 50-80Mb. **Figure 4** shows the imputed and un-imputed GWAS for these three phenotypes in this region, together with histograms of the phenotype data, $r^2$ and missingness metrics. The plot highlights several peaks of association

that harbor a number of genes related to platelet aggregation, adhesion and function (*Igfbp2* and *Igfbp5*[34], *Fn1*[35], *Epha4*[36], *Cps1*[37,38], *Ctla4*[39], *Hspd1*[40]).

An additional association (blue point) in **Figure 3** corresponds to a region associated with the CD25highCD4 phenotype (Proportion of CD4+ cells with high expression of CD25). **Figure 5** shows the imputed and un-imputed GWAS for CD25highCD24 as well as two other related T cell phenotypes that also show increased levels of association (Abs_CD25CD8 (Absolute CD25+CD8+ cell count) and pctDP (Proportion of CD4-CD8- T cells)). The plots show a clear elevation of association in the region around the *Tbx21* (T-bet) gene which plays a key role in T helper cell differentiation [41].

## Discussion

Missing data is a pervasive feature of the statistical analysis of genetic data. Whether it be unobserved genotypes or latent population structure in GWAS studies, partially observed genotypes in low-coverage sequencing studies, or unobserved confounding effects in GWAS and eQTL studies, accurate and efficient methods are needed to infer missing data and can often substantially enhance analysis and interpretation. In this paper, we have proposed a general method to impute missing phenotypes in samples with arbitrary levels of relatedness, population structure and missingness patterns.

While there exists a range of different methods for imputing missing data in the general statistics literature, our method focuses specifically on continuous phenotypes in genetic studies, where there is often known, or measureable, relatedness between samples. Our method leverages this relatedness to partition the phenotypic correlation structure into a genetic and a non-genetic component and to boost imputation accuracy. Using simulated and real data we have shown that our method of imputing missing phenotypes outperforms state-of-the-art methods from the statistics and machine learning literature. In the burgeoning literature of papers on mixed models applied to genetics this is the first approach we are aware of that allows for missing phenotypes.

Key features of our method are (a) boosting signals of association in GWAS when imputation quality is high, (b) not having to discard samples with partially observed phenotypes, (c) a way of assessing imputation performance via our $r^2$ metric, and (d) being able to handle large numbers of phenotypes in a mixed model framework. Our results of applying the method to 140 phenotypes from a rat GWAS study illustrate these key features. However, our results also suggest that imputation will not *always* boost signal, in much the same way the genotype imputation does not always increase levels of association. When imputation quality is demonstrably poor, and missingness is high, then imputation may attenuate the association signal. We recommend filtering phenotype imputation results with the same care and attention as is routine in the analysis of genotype imputation.

The method could be further developed to relax the assumption of normality to directly allow for heavy tailed distributions, or to explicitly allow for binary and categorical traits. However, our simulations have shown that PHENIX remains the currently best performing method in some of these scenarios. In other work (unpublished data; V.I and J.M) we are extending the model to test a SNP for association with multiple phenotypes, using a spike-and-slab mixture prior on effect sizes to allow for only a subset of phenotypes to be associated. Incorporating significant SNPs into our model would likely increase imputation accuracy, especially in model organisms where loci with large effects are common; multi-trait extensions of whole-genome regression models that, intuitively, integrate SNP selection into an LMM-type model[42] could possibly improve accuracy yet further. Higher dimensional datasets, such as '3D' gene expression experiments across multiple samples, genes and tissues[43] also have missing 'phenotypes' which may be reliably imputed to boost signal in downstream analyses.

This paper addresses single imputation (SI) of phenotypes, and ignored uncertainty in these imputed values can, in theory, invalidate subsequent analyses. Multiple imputation (MI), the standard solution, propagates imputation

uncertainty by performing downstream analyses on many imputed datasets, each drawn independently from their posterior. By aggregating results over these multiple datasets, MI delivers valid conclusions for any downstream analysis, regardless the imputation quality[44]. Though drawing from our approximate posterior is not a solution, as VB provably underestimates posterior covariance, it is possible to recover calibrated covariance estimates for the imputed values[45]; doing this computationally efficiently is non-trivial and left to future work. We note that our $r^2$ and missingness metrics dramatically attenuate this shortcoming of SI, as we only analyze phenotypes where imputation uncertainty is smallest; morever, simulations (Supplementary Figure 9) and biologically plausible results (Figures 4 and 5) suggest that SI can uncover novel true positive results in our context.

There is increasing evidence that established loci can affect multiple traits at the same time (pleiotropy)[46] and that this may explain the comorbidity of diseases[47]. It thus seems likely that studies that measure multiple phenotypes, endo-phenotypes and covariates on the same subjects will have to become more common if we are to further elucidate the causal pathways underlying human traits and diseases. Statistical methods that jointly analyze high-dimensional traits and integrate multiple `omics' datasets will be central to this work.

**URLs**

PHENIX : https://mathgen.stats.ox.ac.uk/genetics_software/phenix/phenix.html

TRCMA : http://www.stat.rice.edu/~gallen/software.html

Yeast data :

http://genomics-pubs.princeton.edu/YeastCross_BYxRM/data/cross.Rdata

Wheat data :

http://www.niab.com/pages/id/402/NIAB_MAGIC_population_resources

**Acknowledgments**

J.M acknowledges support from the ERC (Grant no. 617306). A.D. acknowledges support from Wellcome Trust grant [099680/Z/12/Z]. This work was supported

by the Wellcome Trust grant [090532/Z/09/Z]. A.K. acknowledges support from the Royal Society under the Industry Fellowship scheme.

**Author contributions**

A.D, V.I and J.M developed the method. A.D carried out all analysis. J.M and A.D wrote the paper. A.B and R.M provided extensive advice on analysis of the rat GWAS dataset. A.J and U.G provided the NSPHS dataset. N.S provided the UKNBS dataset. A.K provided the chicken dataset and provided advice on analysis. All authors critiqued the manuscript.

**References**

1. Marx, V. Human phenotyping on a population scale. *Nat. Methods* **12,** 711–714 (2015).
2. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* **41,** 1182–1190 (2009).
3. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11,** 407–409 (2014).
4. Huffman, J. E. *et al.* Polymorphisms in B3GAT1, SLC9A9 and MGAT5 are associated with variation within the human plasma N-glycome of 3533 European adults. *Hum. Mol. Genet.* **20,** 5000–5011 (2011).
5. Lauc, G. *et al.* Genomics meets glycomics-the first GWAS study of human N-Glycome identifies HNF1α as a master regulator of plasma protein fucosylation. *PLoS Genet.* **6,** e1001256 (2010).
6. O'Reilly, P. F. *et al.* MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS ONE* **7,** e34861 (2012).
7. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28,** 2540–2542 (2012).
8. Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37,** 710–717 (2005).
9. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62,** 1198–1211 (1998).
10. Abecasis, G. R., Cardon, L. R., Cookson, W. O., Sham, P. C. & Cherny, S. S. Association analysis in a variance components framework. *Genet. Epidemiol.* **21 Suppl 1,** S341–6 (2001).
11. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157,** 1819–1829

(2001).
12. Hai, R. *et al.* Bivariate genome-wide association study suggests that the DARC gene influences lean body mass and age at menarche. *Sci China Life Sci* **55,** 516–520 (2012).
13. Piccolo, S. R. *et al.* Evaluation of genetic risk scores for lipid levels using genome-wide markers in the Framingham Heart Study. *BMC Proc* **3 Suppl 7,** S46 (2009).
14. Choi, Y.-H., Chowdhury, R. & Swaminathan, B. Prediction of hypertension based on the genetic analysis of longitudinal phenotypes: a comparison of different modeling approaches for the binary trait of hypertension. *BMC Proc* **8,** S78 (2014).
15. Scutari, M., Howell, P., Balding, D. J. & Mackay, I. Multiple quantitative trait analysis using bayesian networks. *Genetics* **198,** 129–137 (2014).
16. Park, S. H., Lee, J. Y. & Kim, S. A methodology for multivariate phenotype-based genome-wide association studies to mine pleiotropic genes. *BMC Syst Biol* **5 Suppl 2,** S13 (2011).
17. Cui, X., Sha, Q., Zhang, S. & Chen, H.-S. A combinatorial approach for detecting gene-gene interaction using multiple traits of Genetic Analysis Workshop 16 rheumatoid arthritis data. *BMC Proc* S43 (2009).
18. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46,** 543–550 (2014).
19. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477,** 54–60 (2011).
20. Meuwissen, T. H. E., Odegard, J., Andersen-Ranberg, I. & Grindflek, E. On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genet. Sel. Evol.* **46,** 49 (2014).
21. Schifano, E. D., Li, L., Christiani, D. C. & Lin, X. Genome-wide association analysis for multiple continuous secondary phenotypes. *Am. J. Hum. Genet.* **92,** 744–759 (2013).
22. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–569 (2010).
23. Igl, W., Johansson, A. & Gyllensten, U. The Northern Swedish Population Health Study (NSPHS)--a paradigmatic study in a rural population combining community health and basic research. *Rural Remote Health* **10,** 1363 (2010).
24. Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V. & Kruglyak, L. Finding the sources of missing heritability in a yeast cross. *Nature* **494,** 234–237 (2013).
25. Rat Genome Sequencing and Mapping Consortium *et al.* Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. *Nat. Genet.* **45,** 767–775 (2013).
26. Abdollahi-Arpanahi, R. *et al.* Dissection of additive genetic variability for quantitative traits in chickens using SNP markers. *J. Anim. Breed. Genet.* **131,** 183–193 (2014).
27. Mackay, I. J. *et al.* An eight-parent multiparent advanced generation inter-cross population for winter-sown wheat: creation, properties, and validation. *G3 (Bethesda)* **4,** 1603–1610 (2014).
28. Ferreira, M. A. R. & Purcell, S. M. A multivariate test of association. *Bioinformatics* **25,** 132–133 (2009).

29. Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M., Janss, L. L. & Vermeulen, S. H. A comparison of multivariate genome-wide association methods. *PLoS ONE* **9,** e95923 (2014).
30. Casale, F. P., Rakitsch, B., Lippert, C. & Stegle, O. Efficient set tests for the genetic analysis of correlated traits. *Nat. Methods* **12,** 755–758 (2015).
31. Dahl, A., Hore, V., Iotchkova, V. & Marchini, J. Network inference in matrix-variate Gaussian models with non-independent noise. *arXiv.org* http://arxiv.org/abs/1312.1622v1 (2013).
32. Mott, R., Talbot, C. J., Turri, M. G., Collins, A. C. & Flint, J. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 12649–12654 (2000).
33. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).
34. Hers, I. Insulin-like growth factor-1 potentiates platelet activation via the IRS/PI3Kalpha pathway. *Blood* **110,** 4243–4252 (2007).
35. Cho, J. & Mosher, D. F. Role of fibronectin assembly in platelet thrombus formation. *J. Thromb. Haemost.* **4,** 1461–1469 (2006).
36. Prévost, N. *et al.* Signaling by ephrinB1 and Eph kinases in platelets promotes Rap1 activation, platelet adhesion, and aggregation via effector pathways that do not require phosphorylation of ephrinB1. *Blood* **103,** 1348–1355 (2004).
37. Chen, Y.-R. *et al.* Y-box binding protein-1 down-regulates expression of carbamoyl phosphate synthetase-I by suppressing CCAAT enhancer-binding protein-alpha function in mice. *Gastroenterology* **137,** 330–340 (2009).
38. Shinya, H., Matsuo, N., Takeyama, N. & Tanaka, T. Hyperammonemia inhibits platelet aggregation in rats. *Thromb. Res.* **81,** 195–201 (1996).
39. Gilson, C. R., Patel, S. R. & Zimring, J. C. CTLA4-Ig prevents alloantibody production and BMT rejection in response to platelet transfusions in mice. *Transfusion* **52,** 2209–2219 (2012).
40. Zufferey, A. *et al.* Unraveling modulators of platelet reactivity in cardiovascular patients using omics strategies: Towards a network biology paradigm. *Advances in Integrative Medicine* **1,** 25–37 (2013).
41. Szabo, S. J. *et al.* A Novel Transcription Factor, T-bet, Directs Th1 Lineage Commitment. *Cell* **100,** 655–669 (2000).
42. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9,** e1003264 (2013).
43. GTEx Consortium, Ardlie, K. G. & Dermitzakis, E. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
44. Little, R. J. A. & Rubin, D. B. *Statistical analysis with missing data*. (John Wiley & Sons, Inc., New York, 1987).
45. Giordano, R., Broderick, T. & Jordan, M. Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes. *arXiv.org*, http://arxiv.org/abs/1506.04088v2**,** (2015).
46. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* **7,** e1002254 (2011).
47. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14,** 483–495

(2013).
48. Listgarten, J. *et al.* A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* **29,** 1526–1533 (2013).
49. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44,** 821–824 (2012).
50. Almasy, L., Dyer, T. D. & Blangero, J. Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. *Genet. Epidemiol.* **14,** 953–958 (1997).
51. Mazumder, R., Hastie, T. & Tibshirani, R. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. **11,** 2287–2322 (2010).
52. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17,** 520–525 (2001).
53. Buuren, S. V. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45,** 1–67 (2011).
54. Allen, G. I. & Tibshirani, R. Transposable regularized covariance models with an application to missing data imputation. *Ann Appl Stat* **4,** 764–790 (2010).

## Figure legends

**Figure 1 : Simulation results. Model 1** – scenario simulated using an empirical kinship matrix derived from the human NSPHS study[23]. **Model 2** – scenario simulated using 75 families of 4 sibs. Datasets were simulated at various levels of heritability (x-axis) for the traits. 300 individuals at 15 traits were simulated. 5% of phenotype values were set as missing before imputation. 7 different methods (legend) were applied to impute the missing values. The correlation of the imputed values with the true values is plotted on the y-axis for each method. The lines for TRCMA, MVN and SOFTIMPUTE lie almost exactly on top of each other.

**Figure 2 : Imputation performance in real datasets.** There is one plot for each of the six real datasets. The vertical dotted black line shows the true level of missingness in the dataset. Extra missingness was added to each dataset , and the x-axis shows the amount of missing data in these reduced datasets. The y-axis shows imputation correlation between the imputed missing data and the held out data. The legend denotes the different methods that were applied to the datasets. Not all methods were run on all datasets. TRCMA and MPMM were only run on the human NSPHS and wheat datasets for computational reasons.

**Figure 3 : Missing phenotype imputation in 140 rat GWAS.** The x-axis and y-axis show the –log10(p) for the GWAS on the un-imputed and imputed phenotypes respectively. Each point corresponds to a region in both scans. The dashed black lines denote a conservative threshold of –log10(p)>10 that was applied to highlight associated regions (large points). Points in grey have imputation $r^2$<0.36. Associations with platelet phenotypes on chr 9 and T cell phenotypes on chr 10 are highlighted with red and blue points respectively.

**Figure 4 : Platelet phenotype associations.** GWAS results for un-imputed (blue points) and imputed phenotypes (red points) for three platelet phenotypes (MPC, MPV, PDW) measured in rats, on rat chromosome 9 (50-80Mb). Genes are shown below the plots, with some (named) genes with relevant annotation to platelet function, adhesion and aggregation highlighted in a separate track. Histograms on the right show the distribution of observed (cyan) and imputed (purple) phenotypes, together with missingness and $r^2$ metrics.

**Figure 5 : T cell phenotype associations.** GWAS results for un-imputed (blue points) and imputed phenotypes (red points) for three T cell phenotypes (CD25highCD4, Abs_CD25CD8, pctDP) measured in rats, on rat chromosome 10 (83-89Mb). Genes are shown below the plots, with some (named) genes with relevant annotation to T cell phenotypes highlighted in a separate track. Histograms on the right show the distribution of observed (cyan) and imputed (purple) phenotypes, together with missingness and $r^2$ metrics.

**Tables**

| Method | Description and Properties | References |
|---|---|---|
| PHENIX | Bayesian multivariate mixed model fitted via Variational Bayes | This paper |
| MVN | Multivariate normal model of covariance between traits, fit using an EM algorithm. Ignores genetic covariance between samples. | [44] |
| LMM | Single trait linear mixed model, with estimated BLUP used to impute missing values. Ignores covariance between phenotypes. | [48,49] |
| MPMM | Multiple Phenotype Mixed Model, fit using EM algorithm to only samples without missing data. | [3,50] |
| SOFT-IMPUTE | Low-rank approximation to phenotype matrix via nuclear norm penalty function | [51] |
| kNN | Nearest neighbour imputation | [52] |
| MICE | Multivariate Imputation by Chained Equations | [53] |
| TRCMA | Fits a single matrix normal model to the data by estimating penalized row and column covariances | [54] |

**Table 1 : Brief summary of methods applied to simulated and real datasets**

| Dataset | Number of samples | Number of phenotypes | Missing data (%) | Relatedness Measure | Reference |
|---|---|---|---|---|---|
| Rats | 1,407 | 205 | 15.8 | 0.12 | [25] |

| | | | | | |
|---|---|---|---|---|---|
| Yeast | 1,008 | 46 | 5.2 | 0.10 | [24] |
| Wheat | 720 | 7 | 2.4 | 0.09 | [27] |
| Chickens | 11,575 | 12 | 57.1 | 0.06 | [26] |
| NSPHS | 1,021 | 15 | 0.1 | 0.05 | [23] |
| UKBS | 1,500 | 6 | 14.5 | 0.03 | [2] |

**Table 2 : Summary of real datasets analyzed.** The relatedness measure $\left(\Psi\right)$ is defined in the Online Methods.

## Online methods

Matrix Normal Models

We develop our model using Matrix Normal (MN) distributions [55]. If an $N$x$P$ random matrix $X$ has a Matrix Normal distribution, this is denoted as

$$X \sim MN\left(M,R,C\right)$$

which implies

$$vec\left(X\right) \sim N\left(vec\left(M\right),C \otimes R\right)$$

where *vec(X)* is the column-wise vectorization of $X$, $M$ is the $N$x$P$ mean matrix, $R$ is an $N$x$N$ row covariance matrix, $C$ is a $P$x$P$ column covariance matrix, and $\otimes$ denotes the Kronecker product operator.

A Bayesian Multiple Phenotype Mixed Model

We let $Y$ be an $N$x$P$ matrix of $P$ phenotypes (columns) measured on $N$ individuals (rows). We assume that $Y$ is partially observed and that each phenotype has been de-meaned and variance standardized. A standard Multiple Phenotype Mixed Model (MPMM) has the form

$$Y = U + \varepsilon \quad (1)$$

where $U$ is an $N$x$P$ matrix of random effects and $\varepsilon$ is a $N$x$P$ matrix of residuals and are modeled using Matrix normal distributions as follows

$$U \sim MN\left(0, K, B\right)$$
$$\varepsilon \sim MN\left(0, I_N, E\right) \quad (2)$$

In this model $K$ is the $N$x$N$ kinship matrix between individuals, $B$ is the $P$x$P$ matrix of genetic covariances between phenotypes and $E$ is the $P$x$P$ matrix of residual covariances between phenotypes.

In our Bayesian MPMM (PHENIX), we fit a low-rank model for $U$, such that $U = S\beta$, where

$$S \sim MN\left(0, K, I_P\right)$$
$$\beta \sim MN\left(0, I_P, \tau^{-1} I_P\right) \quad (3)$$

where $\tau$ is a regularization parameter. We use a Wishart prior for the residual precision matrix $E^{-1}$

$$E^{-1} \sim Wi\left(P+5, \frac{1}{4} I_P\right) \quad (4)$$

We fit this model using Variational Bayes (VB) [56], which is an iterative approach of approximating the posterior distribution of the model parameters. We treat missing phenotypes, which we denote as $Y^{(\text{miss})}$, as parameters in the model and infer them jointly with $S, \beta$ and $E$. We impose that the approximate posterior factorizes over the partition $\left\{Y^{(miss)}, S, \beta, E\right\}$. The full details of the VB update equations are given in the **Supplementary Methods.** We let $\tau = 0$ which leads to the least low rank estimate of $U = S\beta$ under our model.

Having fit the model, for each sample with missingness the resulting approximate posterior distribution has the form of a multivariate normal distribution

$$Y_i^{(miss)} \sim N\left(\mu_i, \sigma_i^2 \,|\, Y \setminus Y^{(miss)}\right) \quad (5)$$

We use the posterior mean $\mu_i$ to impute $Y_i^{(\text{miss})}$,.

Other methods

We applied several other methods for imputing missing phenotypes from the statistical genetics, mainstream statistics and machine learning literatures. These methods are summarized briefly in **Table 1.** We provide brief details of each method here and more extensive details in the **Supplementary Methods.**

**MVN** - We assessed the effect of ignoring relatedness between individuals by fitting a simple multivariate normal model of covariance between traits [44]. The model is

$$Y_{i-} \sim N\left(\mu, \sigma^2\right) \tag{6}$$

where $Y_{i-}$ denotes the *i*th row of the phenotype matrix *Y.* We use an expectation-maximization (EM) algorithm that allows for missing phenotypes to fit the model. This method was implemented in R.

**LMM** - To examine the effect of ignoring correlations between traits we applied a single trait linear mixed model (LMM) to each trait separately of the form

$$Y_{-p} \sim N\left(o, \sigma^2_{pg} K + \sigma^2_{pe} I_N\right) \tag{7}$$

where $Y_{-p}$ denotes the *p*th phenotype. Missing phenotypes for each trait were predicted using the BLUP estimate of the random effect. This method was implemented in R.

**MPMM** - We directly fit an MPMM (eqns. 1-2) to only those individuals with completely observed observations, using an EM algorithm (see **Supplementary Methods**) and used the resulting parameter estimates in the model to impute the missing observations. This method was implemented in R.

**TRCMA** - The transposable regularized covariance model (TRCM) approach[54] fits a mean restricted matrix normal model of the form

$$Y \sim MN\left(0, \mu^T 1_P + 1_N \upsilon^T, \Omega^{-1}, \Theta^{-1}\right)$$

where $\Omega$ and $\Theta$ are row and column precision matrices respectively. An EM algorithm fits maximum penalized likelihood estimates, using $L_2$ penalties on both $\Omega$ and $\Theta$, and computes expected values for missing entries. TRCMA is a

one-step approximation to this EM algorithm and was proposed as a computationally tractable alternative[54]. TRCMA is much slower than all other methods we tried in this paper, especially for large *N*. To speed it up, we performed preliminary simulations to determine a small but useful set of regularization parameters to optimize over (5 levels for both the row and column penalties). This method was also run on fewer simulated datasets than the other methods when constructing **Figure 2** due to computational reasons. We used the `R` code from the TRCMA website (see **URLs**) to apply this method.

**SOFTIMPUTE** – there is a large machine learning literature on matrix completion methods [57,58]. We picked a competitive approach [51] which estimates a low-rank approximation to the full matrix of phenotypes via a penalty on the sum of the singular values (or nuclear norm) of the approximation. If *H* is the set of indices of non-missing values in *Y* then the method seeks an estimate, *X*, to the full matrix, *Y*, that minimizes

$$\sum_{i,j \in H} \left( X_{ij} - Y_{ij} \right)^2 + \lambda \left\| X \right\|_*$$

where $\left\| X \right\|_*$ is the nuclear norm of *X*. We used the `R` package `softImpute` to implement this method.

**MICE** – this approach fits regression equations to each phenotype in an iterative algorithm (MICE) and has recently been applied to a metabolite study [18]. We used the `R` package `mice` to implement this method.

**kNN** - We applied a nearest neighbour imputation (kNN) approach which identifies nearest neighbour observations as a basis for prediction [52]. Specifically, if $Y_{ij}$ is a missing phenotype then the *k* nearest phenotypes to phenotype *j* are found, based on all the non-missing values. Then $Y_{ij}$ is predicted by a weighted average of those phenotypes in the *i*th individual. We used the `R` package `impute` to implement this method using the default k =10.

Simulations

We simulated data from the following model

$$Y \sim MN\left(0,K,h^2B\left(\rho\right)\right)+MN\left(0,I,\left(1-h^2\right)E\right)$$

where *K* is the *N*x*N* genetic kinship matrix and $h^2$ is the heritability parameter which we vary between 0 and 1. For the *P*x*P* residual covariance matrix *E* we simulated from a Wishart distribution $Wi\left(P,\frac{1}{P}I_P\right)$, which we then scale to a correlation matrix. For the *P*x*P* genetic covariance matrix *B* we used an AR(1) model with $B\left(\rho\right)_{ij}=\rho^{|i-j|}$. This model produces a range of correlations between traits and is controlled by a single parameter $\rho$. For Figure 1 we used $\rho=0.45$. For **Supplementary Figure 3** we used $\rho=0.275$ and $\rho=0.675$. For the *N*x*N* genetic kinship matrix *K* we used two different models : Model 1 used a subset of the empirical kinship matrix derived from the Northern Sweden Population Health Study (NSPHS) [23]; Model 2 used a kinship structure with independent sets of 4 sibs. We set *N*=300 and *P*=15 for **Figure 1** and *N*=1000 and *P*=50 for **Supplementary Figure 2.** Missing data was added completely at random at the 5% level (**Figure 1**) and 10% (**Supplementary Figure 4**).

Genotype and phenotype data

We analyzed 6 real datasets from 5 different organisms : humans[2,23], rats[25], yeast[24], chickens[26] and wheat[27].

The human data from the UK Blood Services Common Control, collected by the Wellcome Trust Case Control Consortium, include 1,500 individuals with 6 hematological phenotypes (hemoglobin concentration, platelet, white and red blood cell counts, and platelet and red blood cell volume)[2]. DNA samples were genotyped using the Affymetrix 500K GeneChip array. Unassayed genotypes were imputed using IMPUTE2[59] and a 1000 Genomes Project Phase 1 reference panel. We calculated a genetic relatedness matrix (GRM) using code written in R. Following others[1], phenotypes were regressed on the covariates region, age and sex. Extreme outlying measurements were removed to eliminate individuals not representative of normal variation within the population.

The human data from NSPHS[23] include 1,021 individuals with 15 glycans phenotypes (desialylated glycans (DG1-DG13), antennary fucosylated glycans (FUC-A) and core fucosylated glycans (FUC-C)). DNA samples from the NSPHS individuals were genotyped using the Illumina exome chip and either Illumina Infinium HumanHap300v2 (KA06 cohort) or Illumina Omni Express (KA09 cohort) SNP bead microarrays. Unassayed genotypes were imputed using the 1000 Genomes Phase I integrated variant set as the reference panel. Genotype data were imputed with a pre-phasing approach using IMPUTE (version 2.2.2) in the two sub cohorts (KA06 and KA09) separately. We calculated a genetic relatedness matrix (GRM) using GEMMA[3]. We used only those SNPs on either of the two Illumina chips with a minor allele frequency > 1%. Following others[4], phenotypes were regressed on the covariates age and sex and residuals were then quantile normalized. Extreme outlying measurements (those more than three times the interquartile distances away from either the 75th or the 25th percentile values) were removed.

The yeast data[24] was downloaded directly from the web (see **URLs**) and consisted of 1,008 prototrophic haploid segregants from a cross between a laboratory strain and a wine strain of yeast. This dataset was collected via high-coverage sequencing and consists of genotypes at 30,594 SNPs across the genome. There are 46 phenotypes in this dataset and consist of measured growth in multiple conditions, including different temperatures, pHs and carbon sources, as well as addition of metal ions and small molecules[24]. Traits were mean and variance standardized and quantile normalized before analysis. We removed SNPs with MAF < 1% or missingness in > 5% of samples and calculated a GRM using code written in R.

The wheat data[27] was downloaded directly from the web (see **URLs**) and consists of a winter wheat population produced by the UK National Institute of Agricultural Botany (NIAB) comprising 15,877 SNPs for 720 genotypes. Seven traits were measured: yield (YLD), flowering time (FT), height (HT), yellow rust in the glasshouse (YR.GLASS) and in the field (YR.FIELD), Fusarium (FUS), and

mildew (MIL). The population was created using a multiparent advanced generation inter-cross (MAGIC) scheme. Traits were mean and variance standardized and quantile normalized before analysis. We removed SNPs with MAF < 1% or missingness in > 5% of samples and calculated a GRM using code written in R.

The chicken dataset[26] consists of 11,575 samples across 4 full generations of an animal breeding program[26] as part of a collaboration with Aviagen. We used genotypes at 52,679 SNPs. We removed samples that were missing at > 1% of SNPs and SNPs with MAF < 1% or missingness > 5% and calculated a GRM using code written in R. There are 14 traits in this dataset ((BWT) body weight, (LFI) feed intake in females, (AFI) feed intake in males, (WTG) weight gain, (AUS) ultrasound depth, (FL) condition score, (FLMORT) floor mortality , (SLMORT) slat mortality 2, (FPD) foot-pad dermatitis, (HHP) egg production, (EFERT) early fertility , (LFERT) late fertility 2, (EHOF) early hatchability, (LHOF) late hatchability). Each trait was regressed on an appropriate set of covariates, based on experience of the ongoing breeding program. Traits were mean and variance standardized and quantile normalized before analysis.

The GWAS analysis of the rat dataset involves reconstructing the outbred rat genomes as mosaics of 8 founder haplotypes, using the program HAPPY[32]. We obtained the descent probabilities at 24,196 genomic locations based on the Rnor3.4 Rat genome assembly. For the GWAS analysis we obtained the set of pre-processed phenotypes used in the Rat Genome Sequencing and Mapping Consortium paper[25]. In total, we used 317 phenotypes to carry out phenotype imputation. The original study only carried out GWAS for 160 of these traits, deemed to be the most biological relevant traits. We re-analyzed the 140 of these 160 traits that were analyzed using mixed models in the original study. Each trait was analyzed one at a time. For this analysis we used the exact same kinship matrix used in [25]. We also assessed phenotype imputation accuracy on this dataset in Figure 2. We used exactly the 140 phenotypes and the kinship matrix from the GWAS.

When adding additional missing data to the five real datasets, we repeated this

process 100 times for each level of missingness, except for the chicken dataset, which is much larger, where we used 20 simulations. The results are shown in **Figure 2**.

To summarize the overall levels of relatedness in each of the five datasets we calculated the following measure $\left(\Psi\right)$, using the kinship matrix for each dataset

$$\Psi = \sqrt{\sum_{i,j} \left| K_{ij} \right|^2} \Big/ tr\left(K\right)$$

GWAS analysis of outbred rats

To carry out GWAS analysis of the 140 rat phenotypes we used a single-trait mixed model implemented in R. The model consisted of fixed effects that are the estimated founder descent probabilities and covariates, a single random effect with covariance as a scaled kinship and an uncorrelated residual term. This model was fitted at each of the 24,196 genomic locations with descent probabilities. Significance was assessed using an F-test for presence or absence of the descent probabilities in the model. We carried out this analysis twice : before and after phenotype imputation.

Phenotype imputation quality metric ($r^2$)

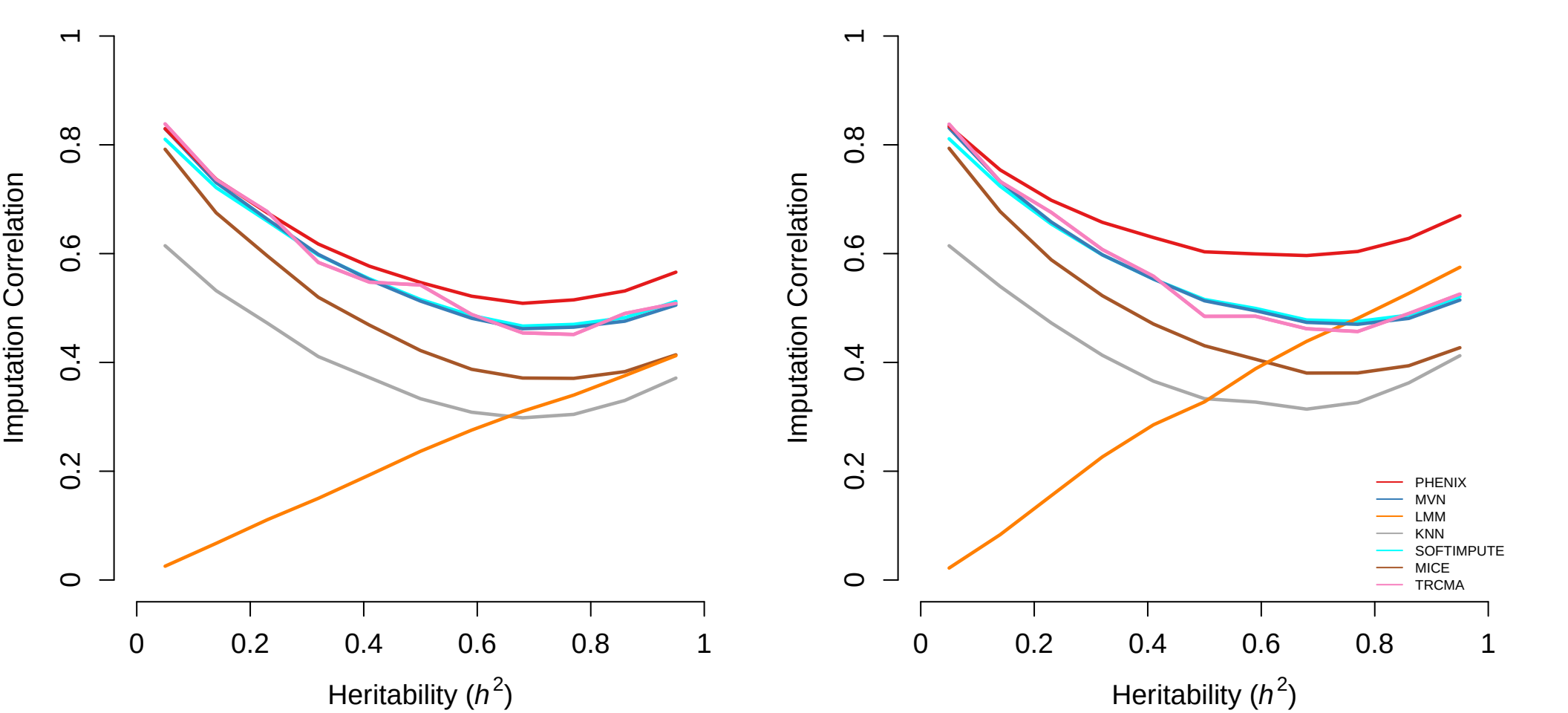
We use real patterns of missing data to simulate extra missing data. We selected a rat at random and then copied its pattern of missing phenotypes to another randomly selected rat. This process continued until an extra 5% of phenotypes had been removed from the dataset. All missing phenotypes were then imputed and the squared correlation ($r^2$) between the imputed values and held out values is calculated. We repeated this process 1,000 times and calculate the mean $r^2$.

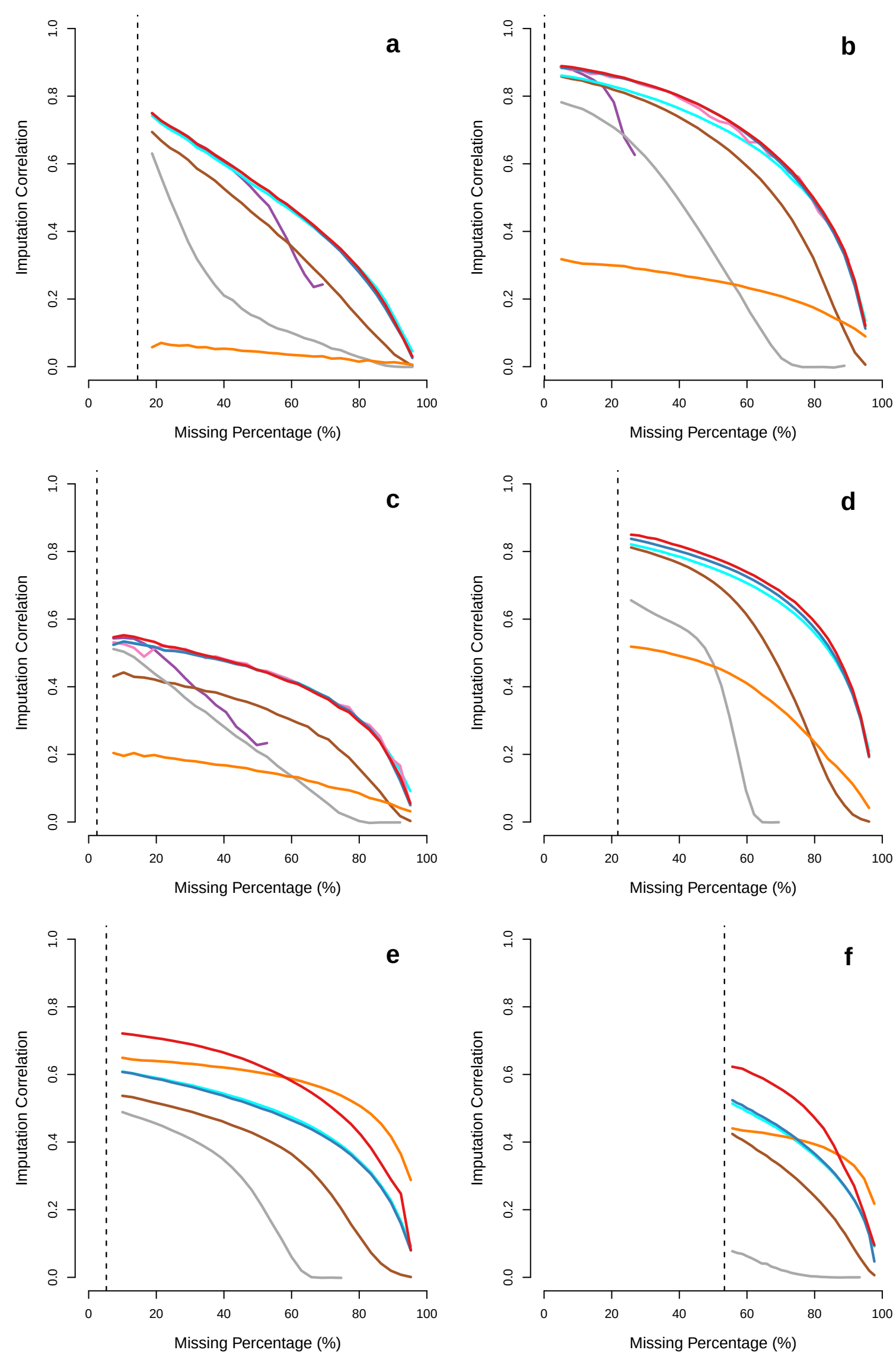

55. Dawid, A. P. Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68,** 265–274 (1981).
56. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. An Introduction to Variational Methods for Graphical Models. *Machine Learning* **37,** 183–233 (1999).
57. Liu, D., Zhou, T., Qian, H., Xu, C. & Zhang, Z. in *Machine Learning and Knowledge Discovery in Databases* **8189,** 210–225 (Springer Berlin

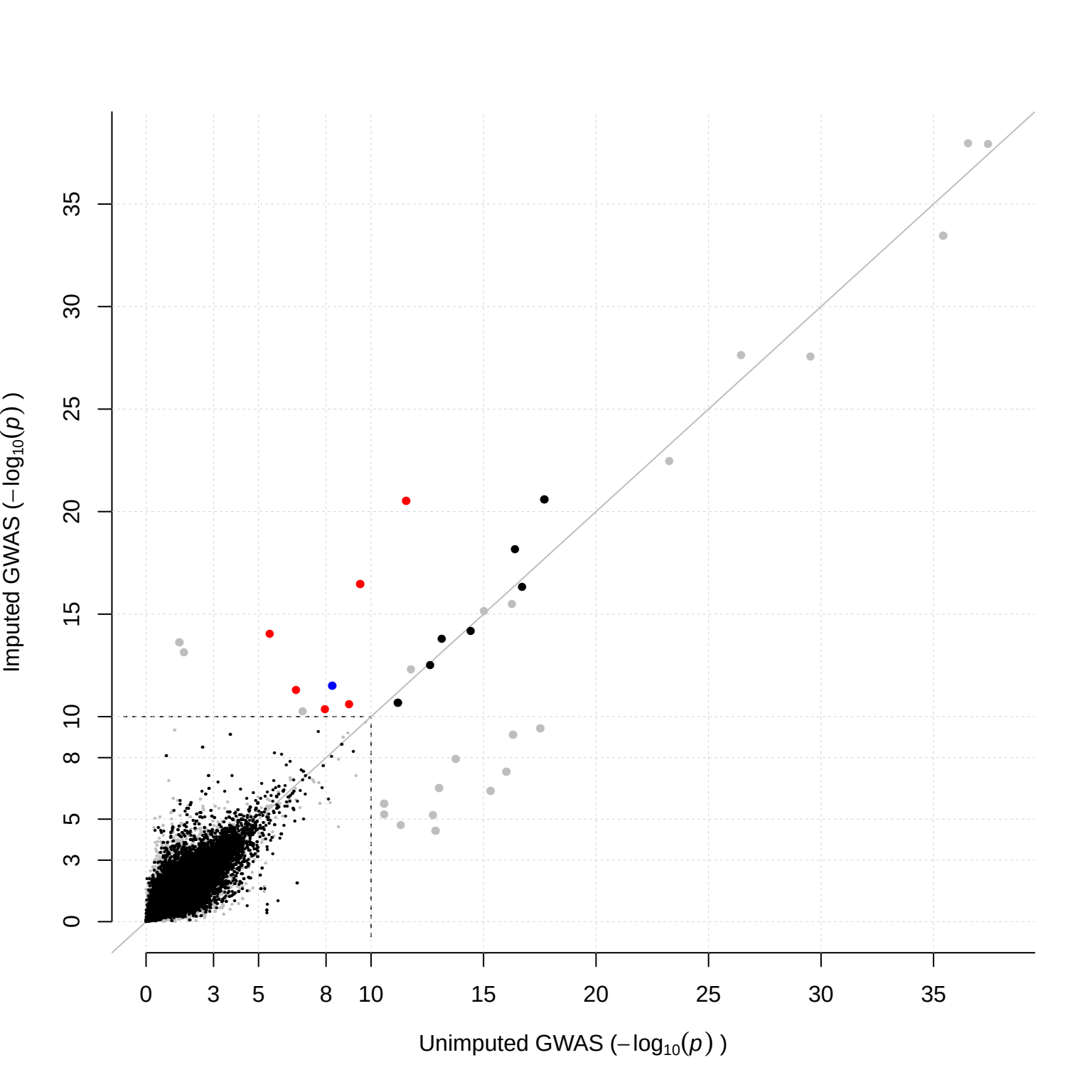
Heidelberg, 2013).
58. Wang, Z. *et al.* Rank-One Matrix Pursuit for Matrix Completion. *ICML* 91–99 (2014).
59. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44,** 955–959 (2012).

**Competing financial interests statement**

AK is an an employee of Aviagen Ltd, a poultry breeding company that supplies broiler breeding stock world wide. AK also holds an Industry Fellowship from the Royal Society and is part-time based in The Roslin Institute.


Imputation Correlation
0
0.2
0.4
0.6
0.8
1
Heritability ($h^2$)
PHENIX
MVN
LMM
KNN
SOFTIMPUTE
MICE
TRCMA


a
b
c
d
e
f
Imputation Correlation
Missing Percentage (%)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100


Imputed GWAS ($-\log_{10}(p)$)
Unimputed GWAS ($-\log_{10}(p)$)
0
3
5
8
10
15
20
25
30
35

−log10(p) for MPC

$r^2$ = 0.461
Miss = 8.1%

−log10(p) for MPV

$r^2$ = 0.448
Miss = 8.0%

−log10(p) for PDW

$r^2$ = 0.476
Miss = 7.3%

0 5 10 15 20

50 55 60 65 70 75 80

Position on rat chromosome 9 (Mb)

Highlighted genes

*Hspd1* *Ctla4* *Cps1* *Erbb4* *Fn1* *Igfbp2* *Igfbp5* *Epha4*

Other named genes

−log10(p) for CD25highCD4

$r^2$ = 0.536
Miss = 59.6%

−log10(p) for Abs_CD25CD8

$r^2$ = 0.606
Miss = 48.4%

−log10(p) for pctDP

$r^2$ = 0.146
Miss = 45.3%

83 84 85 86 87 88 89

Position on rat chromosome 10 (Mb)

Highlighted genes

*Tbx21*

Other named genes

# Multiple phenotype imputation for genetic studies

Andrew Dahl, Valentina Iotchkova, Amelie Baud, Asa Johansson,
Ulf Gyllensten, Nicole Soranzo, Richard Mott, Andreas Kranis,
Jonathan Marchini

February 17, 2016

## Contents

# 1 The PHENIX model

## 1.1 Definitions and notation

The Kronecker product of matrices is denoted by $\otimes$ and the Kronecker sum, $\oplus$, is defined

$$A \oplus B := A \otimes I + I \otimes B$$

For a matrix $X$, we let the lower case $x$ refer to the column-wise vectorization of $X$, written $x = \text{vec}(X)$; similarly, we let $\text{mat}(x) = X$ be the 'inverse' operation (the dimensions being implicitly defined by context). If $M$ is an $NP \times NP$ matrix, we can represent it in terms of $N \times N$ blocks:

$$M = \begin{bmatrix} M_{11} & \dots & M_{1P} \\ \vdots & \ddots & \vdots \\ M_{P1} & \dots & M_{PP} \end{bmatrix}$$

Then the partial trace $tr_P(M)$ is the $P \times P$ matrix of traces of such blocks

$$tr_P(M) = \begin{bmatrix} tr(M_{11}) & \dots & tr(M_{1P}) \\ \vdots & \ddots & \vdots \\ tr(M_{P1}) & \dots & tr(M_{PP}) \end{bmatrix}$$

We write the matrix variate normal with mean $M$, row covariance $R$ and column covariance $C$ as

$$\mathcal{MN}(M, R, C)$$

This is a special case of a multivariate normal as the vectorization of this matrix has mean $\text{vec}(M)$ and covariance $C \otimes R$.

## 1.2 Model description

Let $Y \in \mathbb{R}^{N \times P}$ be a partially observed matrix of $P$ phenotypes measured on $N$ individuals. We assume that the columns of $Y$ have been demeaned and standardized to unit variance. We start with the additive model

$$Y = U + \epsilon \tag{1}$$

where $U$ represents the aggregate genetic contribution to phenotypic variance and $\epsilon$ is idiosyncratic noise. One model we consider uses independent matrix-variate normal distributions for $U$ and $\epsilon$:

$$\begin{array}{rcl} Y & = & U + \epsilon \\ U & \sim & \mathcal{MN}(0, K, B) \\ \epsilon & \sim & \mathcal{MN}(0, I, E) \end{array} \tag{2}$$

$K$ is the kinship matrix between individuals in the sample, which we assume is known from pedigree or genotype data [23, 8, 13, 31, 30, 29]. This model has recently attracted attention in genetics [33, 10, 3, 24] and we refer to it as a multiphenotype mixed model (MPMM).

MPMMs arise as a multiphenotype generalization of the typical univariate linear mixed model (LMM): when $B$ and $E$ are diagonal in (2), the MPMMs reduce to $P$ independent LMMs of the form

$$\begin{array}{rcl} Y_{,p} & = & u_p + \epsilon_p \\ u_p & \overset{\text{ind}}{\sim} & \mathcal{N}(0, B_{pp}K) \\ \epsilon_p & \overset{\text{ind}}{\sim} & \mathcal{N}(0, E_{pp}I) \end{array} \tag{3}$$

Unfortunately, MPMMs can handle only a small number phenotypes, roughly 10 [33]–as $P$ grows, maximum likelihood covariance estimates quickly become both statistically unstable and computationally intractable. Moreover, missing observations are hard to incorporate into MPMMs as the vector of observed phenotypes inherits the matrix normal structure of the full data only if entire rows are missing (see section 2.7). Removing samples with even one missing phenotype [33] thus eliminates the computational aspect of this missing data hurdle, but at the cost of throwing away data; if entries are missing uniformly at random with probability $\theta$, a sample is fully observed with probability $(1-\theta)^P$ and the data waste is exponential in $P$.

To simultaneously address both of these limitations, we develop an alternative multiphenotype generalization of LMMs[1] by assuming an entirely different model for the genetic term $U$. In particular, we use a Bayesian low-rank matrix factorization model for the genetic term $U$. Such low rank models are computationally tractable and, additionally, we believe this rank constraint is often biologically plausible: $U$ will have (approximately) low-rank $M$ when the $P$ observed phenotypes share a simple biological structure that is (mostly) summarized by $M$ latent factors.

Specifically, for $M \leq N, P$, we use the model

$$\begin{array}{rcl} Y|S, \beta, \epsilon & \sim & U + \epsilon \\ U & = & S\beta \\ S & \sim & \mathcal{MN}(0, K, I_M) \\ \beta & \sim & \mathcal{MN}(0, C, B), \\ \epsilon|\Lambda_\epsilon & \sim & \mathcal{MN}\left(0, I, \Lambda_\epsilon^{-1}\right) \\ \Lambda_\epsilon & \sim & \text{Wishart}(e, E) \end{array} \tag{4}$$

If $C$ is allowed to be an arbitrary diagonal matrix[2], then the matrix factorization model in (4) is equivalent to reduced-rank regression in the same sense that MPMM and LMM are equivalent to genome-wide linear regression. For simplicity, we set $C = I_M$, $B = (\tau I_P)^{-1}$, $e = P + 5$ and $E = e^{-1}I_P$ (so that $\mathbb{E}(\Lambda_\epsilon) = I_P$). Though $\tau$ can be tuned by cross-validation, we use the improper $\tau = 0$ by default (see section 1.3.2).

We note that many fast, powerful and robust penalized likelihood methods exist for estimating a spectrally-regularized $U$ in (1), including many focused on imputing missing entries [21, 2, 16, 18]. However, we know of no method that incorporates, or can be easily generalized to incorporate, a non-spherical kinship matrix $K$. But $K$ is the central element of LMMs in genetics (and random effect models generally). Moreover, by comparing to a competitive spectral-regularization algorithm

[1] It actually generalizes a slightly different, Bayesian version of the LMM in (3), where $B_{pp}$ has a scaled $\chi^2$ prior and $E_{pp}$ has an inverse-gamma prior.

[2] Due to scaling and rotation non-identifiablity, $C$ can be assumed diagonal without loss of generality; see, for example, [18].

from the literatue on generic matrix completion [16] (see section 2.6), our simulations and real data analyses suggest incorporating $K$ is always beneficial, and sometimes vital, for imputation accuracy when there is genetic signal.

## 1.3 Variational Bayesian matrix factorization

We use variational Bayes (VB) to approximate the posterior in model (4). In matrix factorization models, VB is an established alternative to MCMC (which can be computationally expensive) and maximum *a posteriori* [22, 7, 12] (which can suffer from over-fitting). Moreover, VB matrix factorization has known theoretical properties in special cases [18] (see section 1.3.1). Our implementation iteratively updates approximate posteriors on $S$, $\beta$, $\Lambda_\epsilon$ and $Y^m$, the missing entries of $Y$, assuming that these parameters are independent in the posterior. Though this independence assumption does not hold and is potentially problematic [22], it simplifies computation while hopefully retaining much of the exact problem's structure.

Specifically, we require $Q$, the variational approximation to the posterior, to factorize over the partition $\{S, \beta, \Lambda_\epsilon, Y^m\}$ of the parameter space:

$$Q(Y^m, S, \beta, \Lambda_\epsilon | Y \setminus Y^m) := Q_Y(Y^m) Q_S(S) Q_\beta(\beta) Q_\epsilon(\Lambda_\epsilon)$$

The goal is then to find $Q$'s that best approximate the posterior (in Kullback-Leibler divergence). Defining $m_i$ as the missing phenotypes for sample $i$, section 1.4 shows that the $Q$'s belong to simple parametric families:

$$\begin{aligned} Q(Y_{i,m_i}) &\sim \mathcal{N}\left(\mu^{Y_i}, \Sigma^{Y_i}\right) \\ Q(\mathrm{vec}\,(S)) &\sim \mathcal{N}\left(\mu_s, \Lambda_s^{-1}\right) \\ Q(\mathrm{vec}\,(\beta)) &\sim \mathcal{N}\left(\mu_b, \Lambda_b^{-1}\right) \\ Q(\Lambda_\epsilon) &\sim \mathrm{W}\left(e', \frac{1}{e'}\Omega\right) \end{aligned}$$

The problem of optimizing the $Q$'s thus reduces to finding optimal variational parameters for the above approximate marginals.

This minimization is performed by iterating through conditional modes, optimizing each approximate marginal given the others (see Section 1.4.1). Because the conditional optimizers have analytic expressions, this hill-climbing is fast. Unfortunately, this coordinate ascent need not reach a global optimum as our variational objective is non-convex (in addition to the rotation ambiguity in the product $S\beta$, which is inconsequential since we never jointly update $S$ and $\beta$) [7]. Nonetheless, we have not found this problematic in our setting: maybe this is because we initialize at full rank $S\beta$ and allow the fitted rank to converge from above (see 1.3.1 and 1.3.2); maybe it is because we initialize with another method (MVN); maybe it is because we update all of $S$ or $\beta$ at once, avoiding the typical practice of conditionally updating each component given the others.

As written, the approximate marginals for $S$ and $\beta$ depend on very large precision matrices–$\Lambda_s$ and $\Lambda_b$–that induce $O(M^3(N^3 + P^3))$ computations. Though these matrices are not Kronecker products–and so $S$ and $\beta$ are not matrix normal, even in our variational approximation to the posterior–they do have a simple structure that admits much faster computations. If $N_m$ is the number of unique missingness patterns among samples, our algorithm costs $O(N_m P^3 + NP^2 + N^2 M)$ for each VB iteration; additionally, we perform a one-off, full-rank eigendecomposition of $K$ at $O(N^3)$.

### 1.3.1 Properties of a special case

The globally optimal VB matrix factorization parameters have analytic expressions when $Y$ is fully observed and covariances are spherical ($\Lambda_\epsilon = I_P$ and $K = I_N$) [18]. As those authors note, these equations do not easily generalize either to missing data or to non-spherical priors, and this result is not directly useful for us.

Nonetheless, these analytic solutions reveal a surprising property of VB matrix factorization: $\hat{U}$, the expected $U$ under the approximate posterior, may have rank strictly less than $M$, the *a priori* maximum rank of $U$ and the almost-sure rank of $U$ under both the prior and the (exact) posterior. This is because the singular values of $\hat{U}$ are, roughly, the soft-threshholded singular values of $Y$[3]. As $\tau$ controls the magnitude of this soft-threshholing, the search over $\tau$ can replace the search over $M$, much as (convex) lasso relaxes (non-convex) subset search for regression. In fact, reasonable conditions guarantee that optimizing $\tau$ is enough to recover the correct rank of $U$ [19].

Though these automatic rank selection properties have not been proven in our context, we assume that analogues apply as we have consistently observed that our model fits low-rank $\hat{U}$. Specifically, we assume that the automatic rank determination is reliable, so we always set $M = \min(N, P)$–a computational impossibility for truly large $P$–and allow the algorithm to decide the rank of the putatively low-rank component through $\tau$.

### 1.3.2 Choosing the regularization parameter $\tau$

Surprisingly, even when $\tau = 0$ and the prior on $\beta$ is flat, the implied prior on the product $U = S\beta$ is non-flat and shrinks the singular values of $U$ to zero (see section 4). Nonetheless, increasing $\tau$ increases regularization, motivating $\tau = 0$ as a widely applicable default, as this value is optimal for all datasets where even this minimal amount of shrinkage is too much; for example, cross-validation chose $\tau = 0$ of its own accord in the NSPHS data set. In all analyses in the paper we have only used $\tau = 0$.

## 1.4 Details of the PHENIX algorithm

### 1.4.1 Variational Bayes overview

VB aims to approximate a complicated posterior distribution $P(\theta|D)$, where $D$ is the data and $\theta \in \Theta$ are the model parameters, by a function $Q(\theta)$ chosen from a class of simple functions, $\mathcal{Q}$. Once found, exact properties of the approximate posterior, $Q$, can be used to approximate properties of the exact posterior, $P(\cdot|D)$, such as parameter means and covariances and marginal likelihoods.

For any approximate posterior $Q$, the true log marginal likelihood can be written as

$$\log P(D) \quad = \quad F(Q) + D_{KL}(Q||P(\cdot|D)) \tag{5}$$

where $D_{KL}$ is the Kullback-Liebler divergence and $F(Q) = \int \log \left[\frac{P(\theta,D)}{Q}\right] dQ(\theta)$. We choose $Q \in \mathcal{Q}$ to minimize $D_{KL}$ which, since the marginal likelihood $P(D)$ does not depend on $Q$, is equivalent to maximizing $F(Q)$. Moreover, since $D_{KL}$ is non-negative, $F(Q)$ lower-bounds, and approximates, the log marginal likelihood.

[3]This is made formal in [18]; see also [9], which relates the variational Bayesian matrix factorization objective to nuclear norm regularization and, thus, to the matrix completion methods in [16, 2, 21]

Mean field approximations are one way to specify $\mathcal{Q}$, which require that each $Q \in \mathcal{Q}$ factorizes over some partition of $\Theta$:

$$Q \in \mathcal{Q} \iff Q(\theta) = \prod_i Q_i(\theta_i) \quad \forall \theta \in \Theta$$

With this mean field assumption, it is natural to iteratively optimize one coordinate of $Q$ given the others:

$$Q_i \leftarrow \arg\max_{Q_i'} F(Q_i', Q_{-i}) \tag{6}$$

Since we are minimizing $D_{KL}$, these updates take a particularly simple form:

$$\log Q_i \leftarrow \arg\max_{Q_i} F(Q_i, Q_{-i}) \equiv \mathbb{E}_{\theta_{-i} \sim Q_{-i}} \Big( \log P(D, \theta) \Big) \tag{7}$$

The precise form of each $Q_i$ will depend on the likelihood and priors, and one key feature is that the $Q_i$ are not chosen in advance but rather chosen to minimize Kullback-Leibler divergence from the posterior. Nonetheless, the usefulness of VB typically relies on each $Q_i$ reducing to a tractable parametric form, which we index by variational parameters $\tilde{\theta}_i$. With this simplification, the coordinate ascent problem (6), which in general optimizes $Q_i$ over a function space, reduces to optimizing $\tilde{\theta}_i$.

Since we require $Q$ to factorize over the parameter partition $\{S, \beta, Y^m, \Lambda_\epsilon\}$, our mean field algorithm iteratively updates $Q_S$, $Q_\beta$, $Q_\epsilon$ and $Q_Y$. Below, we use (7) to derive these updates.

### 1.4.2 The parametric forms of the approximate posterior marginals

$\boldsymbol{Y}$ : $Q_{Y_{i,m_i}} \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu^Y_{i,m_i}, \Sigma^{Y_i}\right)$

$$\begin{aligned}
-2 \log Q_{Y^m} &\equiv -2\mathbb{E}_{-Y^m} \left( \log P(Y | S, \beta, \Lambda_\epsilon) \right) \\
&\equiv \mathbb{E}_{-Y^m} \left( \operatorname{tr} \left( (Y - S\beta) \Lambda_\epsilon (Y - S\beta)^T \right) \right) \\
&\equiv \operatorname{tr} \left( (Y - \mathbb{E}(S\beta)) \mathbb{E}(\Lambda_\epsilon) (Y - \mathbb{E}(S\beta))^T \right) \implies \\
Q_{Y_i} &\overset{\text{ind}}{\sim} \mathcal{N}\left( (\mu_S \mu_\beta)_{i,}, \Omega^{-1} \right)
\end{aligned}$$

where $\mu_S$, $\mu_\beta$ and $\Omega$ are moments of the other marginals and defined by their respective updates (see below). The distribution of $Y^m | Y^o$ follows from this unconditional distribution:

$$\begin{aligned}
Y_{i,m_i} | Y_{i,o_i} &\overset{\text{ind}}{\sim} \mathcal{N}\left(\mu^Y_{i,m_i}, \Sigma^{Y_i}\right) \\
\mu^Y_{i,m_i} &= (\mu_S \mu_\beta)_{i,m_i} + (\Omega^{-1})_{m_i,o_i} \left(\Omega^{-1}\right)_{o_i,o_i} (Y_{i,o_i} - (\mu_s \mu_\beta)_{i,o_i})^T \qquad (8) \\
\Sigma^{Y_i} &= (\Omega_{m_i,m_i})^{-1} \qquad (9)
\end{aligned}$$

Updating $\mu^Y_{i,m_i}$ and $\Sigma^{Y_i}$ for each $i$ costs $O(NP^3)$. But, since the $O(P^3)$ operations for each $i$ depend on $i$ only through $o_i$, the complexity can be reduced to $O(NP^2 + N_m P^3)$, where $N_m$ is the number of unique trait missingness patterns among the $N$ samples. In real datasets, where experimental and observational constraints often induce highly structured missingness patterns, $N_m$ is often much smaller than $N$: for example, in the chicken data, $N = 11,575$ but $N_m = 36$.

$$\boldsymbol{\beta}: Q_{\mathbf{vec}(\beta)} \sim \mathcal{N}\left(\mu_b, \Lambda_b^{-1}\right)$$

$$\begin{aligned}
-2\log Q_\beta &\equiv -2\mathbb{E}_{-\beta}\left(\log P(Y|S,\beta,D,\Lambda_\epsilon) + \log P(\beta)\right) \\
&\equiv \mathbb{E}_{-\beta}\left(||(Y - S\beta)\Lambda_\epsilon^{1/2}||_F^2 + \tau||\beta||_F^2\right) \\
&\equiv \mathrm{tr}\left(\beta\mathbb{E}\left(\Lambda_\epsilon\right)\beta^T\mathbb{E}\left(S^TS\right)\right) - 2\mathrm{tr}\left(\beta\mathbb{E}\left(\Lambda_\epsilon Y^TS\right)\right) + \tau\,\mathrm{tr}\left(\beta\beta^T\right) \\
&\equiv \mathrm{vec}\left(\beta\right)^T\left[\mathbb{E}\left(\Lambda_\epsilon\right)\otimes\mathbb{E}\left(S^TS\right) + \tau I\right]\mathrm{vec}\left(\beta\right) - 2\mathrm{vec}\left(\beta\right)^T\mathrm{vec}\left(\mathbb{E}\left(S^TY\Lambda_\epsilon\right)\right) \implies \\
\mathrm{vec}\left(\beta\right) &\sim \mathcal{N}\left(\mu_b, \Lambda_b^{-1}\right)
\end{aligned}$$

giving the updates

$$\Lambda_b = \Omega_\beta \otimes V_S + \tau I \qquad \text{(implicit)}$$
$$\mu_b = \Lambda_b^{-1}\mathrm{vec}\left(\mu_S^T\mu_Y\Omega_\beta\right) \qquad (10)$$
$$\Omega_\beta = \Omega \qquad (11)$$
$$V_S = \mu_S^T\mu_S + \mathrm{tr}_P\left(\Lambda_s^{-1}\right) \qquad (12)$$

Using lemma 2, (10) can be computed in $O(P^3 + MNP)$ rather than $O(M^3P^3 + MNP)$. Similarly, using lemma 1, (12) can be found in $O(M^3 + NM^2)$ rather than $O(N^3M^3)$. In both cases, explicitly forming $\Lambda_b$ is unnecessary; because $\Lambda_b$ is a function of a specific $\Omega$, not whatever $\Omega$ has become since last updating $Q_\beta$, we perform (11) so we can at all times evaluate terms involving $\Lambda_b$.

$$\boldsymbol{S}: Q_{\mathbf{vec}(S)} \sim \mathcal{N}\left(\mu_s, \Lambda_s^{-1}\right)$$

$$\begin{aligned}
-2\log Q_{-S} &\equiv -2\mathbb{E}_{-S}\left(\log P(Y|S,\beta,D,\Lambda_\epsilon) + \log P(S)\right) \\
&\equiv \mathbb{E}_{-S}\left(||(Y - S\beta)\Lambda_\epsilon^{1/2}||_F^2 + ||K^{-1/2}S||_F^2\right) \\
&\equiv \mathrm{tr}\left(S\mathbb{E}\left(\beta\Lambda_\epsilon\beta^T\right)S^T\right) - 2\mathrm{tr}\left(S\mathbb{E}\left(\beta\Lambda_\epsilon Y^T\right)\right) + \mathrm{tr}\left(S^TK^{-1}S\right) \\
&\equiv \mathrm{vec}\left(S\right)^T\left(\mathbb{E}\left(\beta\Lambda_\epsilon\beta^T\right)\otimes I + I\otimes K^{-1}\right)\mathrm{vec}\left(S\right) - 2\mathrm{vec}\left(S\right)^T\mathrm{vec}\left(\mathbb{E}\left(Y\Lambda_\epsilon\beta^T\right)\right) \implies \\
\mathrm{vec}\left(S\right) &\sim \mathcal{N}\left(\mu_s, \Lambda_s^{-1}\right)
\end{aligned}$$

where

$$\Lambda_s = V_\beta \oplus K^{-1} \qquad \text{(implicit)}$$
$$\mu_s = \Lambda_s^{-1}\mathrm{vec}\left(\mu_Y\Omega\mu_\beta^T\right) \qquad (13)$$
$$V_\beta = \mu_\beta\Omega\mu_\beta^T + \mathrm{tr}_P\left(\left(\Omega\otimes I\right)\Lambda_b^{-1}\right) \qquad (14)$$

Since only explicitly evaluated parameters depend on $\Omega$, there is no need to store a copy.

Unfortunately, $\mathrm{tr}_P\left(\left(\Omega\otimes I\right)\Lambda_b^{-1}\right)$ does not generally simplify as $\Omega \neq \Omega_\beta$ in general. However, I ensure $Q_\beta$ was updated more recently than $Q_\epsilon$ when updating $Q_S$, and so $\Omega = \Omega_\beta$ and

$$\mathrm{tr}_P\left(\left(\Omega\otimes I\right)\Lambda_b^{-1}\right) = \mathrm{tr}_P\left(\left[\left(\tau\Omega^{-1}\right)\otimes V_S\right]^{-1}\right)$$

With this simplification, lemma 2 computes (13) in $O(N^2M + P^2M)$ instead of $O(N^3M^3 + P^2M)$, lemma 1 computes (14) in $O(P^3)$ rather than $O(M^3P^3)$ and $\Lambda_s$ need not be evaluated.

Equation (13) is the reason our method has $O(N^2M)$ iterations while most mixed models only have one $O(N^2P)$ step: typical mixed models assume $Y$ is complete and so the problematic step, whitening $Y$ (or, in our case, $\mu_Y$), only needs to be performed once[4].

Equation (13) is also where low-rank kinship models pay off: if $\text{rk}(K) = R$, the cost of this step becomes $O(NRM + P^2M)$ and the overall complexity drops from $O(N_mP^3 + NP^2 + N^2M)$ to $O(N_mP^3 + NP^2 + NRM)$. Though this change will be crucial for small $P$, huge $N$–where $N$ is, say, tens or hundreds of thousands and $P$ is, say, tens–it is unlikely to matter much in our currently studied applications; a similar logic applies to the one-off, low-rank eigendecomposition of $K$, which can be sped up to $O(RN^2)$.

$\mathbf{\Lambda_\epsilon} : Q_\epsilon \sim \mathbf{W}\left(e', \frac{1}{e'}\Omega\right)$

Define $\tilde{\Sigma}^{Y_i} \in \mathbb{R}^{P\times P}$ by padding $\Sigma^{Y_i} \in \mathbb{R}^{m_i\times m_i}$ with 0s in the natural way. Then

$$\begin{aligned}
\Omega_0 &:= \mathbb{E}\left((Y - S\beta)^T(Y - S\beta)\right) \\
&= \mathbb{E}\left((Y - \mathbb{E}\left(S\beta\right))^T(Y - \mathbb{E}\left(S\beta\right))\right) + \mathbb{E}\left(\left(S\beta - \mathbb{E}\left(S\beta\right)\right)^T\left(S\beta - \mathbb{E}\left(S\beta\right)\right)\right) \\
&= (\mu_Y - \mu_S\mu_\beta)^T(\mu_Y - \mu_S\mu_\beta) + \sum_n \tilde{\Sigma}^{Y_n} \qquad (16) \\
&\quad + \mu_\beta^T \text{tr}_P\left(\Lambda_s^{-1}\right)\mu_\beta + \text{tr}_P\left((I \otimes \left[\mu_S^T\mu_S + \text{tr}_P\left(\Lambda_s^{-1}\right)\right])\Lambda_b^{-1}\right) \qquad (17)
\end{aligned}$$

If $Q_\beta$ has been updated more recently than $Q_S$, $V_S = \mu_S^T\mu_S + \text{tr}_P\left(\Lambda_s^{-1}\right)$ and then

$$\text{tr}_P\left((I \otimes \left[\text{tr}_P\left(\Lambda_s^{-1}\right) + \mu_S^T\mu_S\right])\Lambda_b^{-1}\right) = \text{tr}_P\left(\left[\Omega_\beta \oplus (\tau V_S^{-1})\right]^{-1}\right)$$

Now the $\text{tr}_P(\cdot)$ terms are inverse Kronecker sums and so, by lemma 1, (17) costs $O(P^3 + NM)$ to evalate; (16) costs $O(NP^2)$ as written.

[4]We could save some computation by storing a whitened version of the observed parts of $Y$. Let $Y^0_{ij} = Y_{ij}$ if observed, $Y^0_{ij} = 0$ otherwise. Then store

$$Y' = Q^TY^0$$

where $Q$ are the eigenvectors of $K$. Then at each iteration, $Q^T\mu_Y$ can be computed by

$$Q^T\mu_Y = Y' + Q^TY^1 \qquad (15)$$

where $Y^1_{ij} = 0$ if $Y_{ij}$ is observed and $Y^1_{ij} = \mu^Y_{ij}$ otherwise. Since $Y^1$ has only $n_{miss}$ nonzero entries, the multiplication in (15) is $O(Nn_{miss})$, which may be substantially cheaper than $O(N^2M)$ in some applications. Nonetheless, $n_{miss}$ will almost always be $O(NP)$ and so the $O(Nn_{miss})$ cost is only superficially linear in $N$; in fact, this cost may be greater than $O(N^2M)$ when $M << P$.

Letting $e' = e + N$, it then follows that

$$\begin{aligned}
\log Q_\epsilon(\Lambda_\epsilon) &\equiv \mathbb{E}_{-\Lambda_\epsilon}\left(\log P(Y|S,\beta,\Lambda_\epsilon) + \log P(\Lambda_\epsilon)\right) \\
&\equiv \mathbb{E}_{-\Lambda_\epsilon}\left(-\frac{1}{2}\mathrm{tr}\left((Y-S\beta)\Lambda_\epsilon(Y-S\beta)^T\right) + \frac{N}{2}\log|\Lambda_\epsilon|\right) + \left(\frac{e-P-1}{2}\log|\Lambda_\epsilon| - \frac{1}{2}\mathrm{tr}\left(E^{-1}\Lambda_\epsilon\right)\right) \\
&\equiv -\frac{1}{2}\mathrm{tr}\left(\Lambda_\epsilon\left(\mathbb{E}\left((Y-S\beta)^T(Y-S\beta)\right) + E^{-1}\right)\right) + \frac{N+e-P-1}{2}\log|\Lambda_\epsilon| \implies \\
Q_\epsilon &\sim \mathrm{Wi}\left(e', \frac{1}{e'}\Omega\right)
\end{aligned}$$

where

$$\Omega := e'\left(\Omega_0 + E^{-1}\right)^{-1}$$

### 1.4.3 The marginal likelihood lower bound

We assess convergence by monitoring relative change in the marginal likelihood lower bound ($F(Q)$ in (5)); by default, we terminate once either 1,000 iterations have been performed or the relative change in $F(Q)$ is less than $10^{-8}$.

At the current set of variational parameters $\tilde{\theta}$, the variational posterior is $Q_{\tilde{\theta}}$–$Q$ for short–and the marginal likelihood lower bound is

$$\begin{aligned}
F(Q) &= \mathbb{E}_{\theta\sim Q}\left(\log P(Y^o,\theta) - \log Q(\theta)\right) \\
&= \mathbb{E}_Q\left(\log P(Y^o, Y^m, \beta, S, \Lambda_\epsilon) - \log Q(Y^m,\beta,S,\Lambda_\epsilon)\right) \\
&= \mathbb{E}_Q\left(\log P(Y|\beta,S,\Lambda_\epsilon) + \log P(\Lambda_\epsilon) - \log Q_Y(Y^m) - \log Q_\epsilon(\Lambda_\epsilon)\right) && (18) \\
&\quad + \mathbb{E}_Q\left(\log P(\beta) - \log Q_\beta(\beta)\right) && (19) \\
&\quad + \mathbb{E}_Q\left(\log P(S) - \log Q_S(S)\right) && (20)
\end{aligned}$$

We now compute each part:

$$\begin{aligned}
(18) &= 2\mathbb{E}_Q\left(\log P(Y|\beta,S,\Lambda_\epsilon) + \log P(\Lambda_\epsilon) - \log Q_Y(Y^m) - \log Q_\epsilon(\Lambda_\epsilon)\right) \\
&\equiv \mathbb{E}_Q\Big(N\log|\Lambda_\epsilon| - \left|\left|Y-S\beta\right|\right|^2_{\Lambda_\epsilon} + (e-P-1)\log|\Lambda_\epsilon| - \mathrm{tr}\left(E^{-1}\Lambda_\epsilon\right) \\
&\qquad - \sum_n\left(-\log|\Sigma^{Y_n}| - (Y_{nm}-\mu^Y_{nm}){\Sigma^{Y_n}}^{-1}(Y_{nm}-\mu^Y_{nm})^T\right) \\
&\qquad - \left(-e'\log|\Omega| + (e'-P-1)\log|\Lambda_\epsilon| - \mathrm{tr}\left(\Omega^{-1}\Lambda_\epsilon\right)\right) \\
&\equiv \mathbb{E}_Q\left(-\mathrm{tr}\left(\left[(Y-S\beta)^T(Y-S\beta) + E^{-1}\right]\Lambda_\epsilon\right)\right) + \sum_n \log|\Sigma^{Y_n}| + e'\log|\Omega| \\
&\equiv -\mathrm{tr}\left(\left[\Omega_0' + E^{-1}\right]\Omega\right) + \sum_n \log|\Sigma^{Y_n}| + e'\log|\Omega|
\end{aligned}$$

where $\Omega_0'$ is an up-to-date version of the $\Omega_0$ defined above; in particular, I ensure $\Omega$ was the last

update, so $\text{tr}\left(\left[\Omega_0' + E^{-1}\right]\Omega\right) = e' \equiv 0$.

$$\begin{aligned}
(19) &= 2\mathbb{E}_Q\left(\log P(\beta) - \log Q_\beta(\beta)\right) \\
&= \mathbb{E}_Q\left(-\tau||\beta||_F^2 - \log|\Lambda_b| + (b-\mu_b)^T\Lambda_b(b-\mu_b)\right) \\
&\equiv -\tau\mathbb{E}_Q\left(||\beta||_F^2\right) - \log|\Lambda_b| \\
&\equiv -\tau\left(||\mu_\beta||_F^2 + \text{tr}\left(\Lambda_b^{-1}\right)\right) - \log|\Lambda_b| \\
(20) &= 2\mathbb{E}_Q\left(\log P(S) - \log Q_S(S)\right) \\
&= \mathbb{E}_Q\left(-||S||_{K^{-1}} - \log|\Lambda_s| + (s-\mu_s)^T\Lambda_s(s-\mu_s)\right) \\
&= -\text{tr}\left(\mathbb{E}_Q\left(SS^T\right)K^{-1}\right) - \log|\Lambda_s| \\
&= -\text{tr}\left(\mu_S^T K^{-1}\mu_S\right) - \text{tr}\left((I\otimes K^{-1})\Lambda_s^{-1}\right) - \log|\Lambda_s|
\end{aligned}$$

Altogether, the marginal likelihood lower bound is

$$\sum_n \log|\Sigma^{Y_n}| + e'\log|\Omega| - \tau\left(||\mu_\beta||_F^2 + \text{tr}\left(\Lambda_b^{-1}\right)\right) - \log|\Lambda_b| - \text{tr}\left(\mu_S^T K^{-1}\mu_S\right) - \text{tr}\left((I\otimes K^{-1})\Lambda_s^{-1}\right) - \log|\Lambda_s|$$

All terms can be computed in $O(N_m P^3 + NP^2)$, again assuming updates have been performed in the order necessary for computations to simplify.

# 2 Other methods for imputing missing phenotypes

## 2.1 MVN: an EM algorithm assuming unrelated samples

Rows of $Y$ are not independent in the presence of genetic relatedness between samples due to either population structure or causal genes. Nonetheless, a simple EM algorithm can be derived assuming

$$Y_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$$

The resulting EM algorithm infers $\Sigma$ in an M-step and, among other things, the missing entries of $Y$ in an E-step [14]. As this method ignores correlation across samples, it should do well when there is either little relatedness or little heritability.

### Derivation

Given a current parameter estimate $\hat{\Sigma}$, the expected log likelihood is

$$Q(\Sigma|\hat{\Sigma}) \equiv -N\log|\Sigma| - \sum_{n=1}^N \text{tr}\left(\Sigma^{-1}\mathbb{E}_{Y^m|Y^o,\hat{\Sigma}}\left(Y_n Y_n^T\right)\right)$$

where $m$ and $o$ are missing and observed entries, respectively. Letting $m_n$ and $o_n$ be the missing and observed entries of sample $n$, respectively, define $\hat{Y}$, the implicitly imputed phenotypes, by

$$\hat{Y}_{no_n} = Y_{no_n}, \quad \hat{Y}_{nm_n} = \mathbb{E}\left(Y_{nm_n}|Y_{no_n}, \hat{\Sigma}\right) = \hat{\Sigma}_{m_n o_n}\hat{\Sigma}_{o_n o_n}^{-1} Y_{no_n}$$

Now define the expected sample covariance

$$S := \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{Y^m|Y^o,\hat{\Sigma}}\left(Y_n Y_n^T\right)$$

where

$$\begin{aligned}\mathbb{E}_{Y^m|Y^o,\hat{\Sigma}}\left(Y_n Y_n^T\right)_{i,j} &= \left(\hat{Y}_n \hat{Y}_n^T\right)_{i,j} + \mathrm{Cov}\left(Y_{ni}, Y_{nj}|Y^o, \hat{\Sigma}\right) \\ &= \left(\hat{Y}_n \hat{Y}_n^T\right)_{i,j} + \mathbb{I}\{i,j \in m_n\}\,\Sigma_{ij}^{(n)} \\ \text{where } \Sigma_{ij}^{(n)} &:= \Sigma_{ij} - \Sigma_{i,o_n}\left(\Sigma_{o_n,o_n}\right)^{-1}\Sigma_{o_n,j}\end{aligned}$$

so that

$$Q(\Sigma|\hat{\Sigma}) \equiv -N\log|\Sigma| - \mathrm{tr}\left(\Sigma^{-1}S\right) \implies \Sigma^{(t+1)} = S$$

## 2.2 LMM: univariate linear mixed models

For each phenotype independently, we run a linear mixed model (LMM) on the observed samples to find the MLE variance components ($B_{pp}$ and $E_{pp}$ in terms of (2)) and then, using these estimates, impute missing samples to their conditional expectations, or BLUPs:

$$\hat{Y}_{m_p,p} := B_{pp}K_{m_p,o_p}\left(B_{pp}K_{o_p,o_p} + E_{pp}I\right)^{-1}Y_{o_p,p}$$

We use the computational trick from [25, 13] to expedite variance component estimation; that is, we first rotate $Y$ by the eigenvectors of $K$ so that the entries of the resulting vector are independent.

## 2.3 TRCMA: transposable regularized covariance model

The transposable regularized covariance model of [1] (TRCM) uses a mean-restricted matrix normal:

$$Y \sim \mathcal{MN}\left(1_N\mu^T + \nu 1_P^T, R, C\right)$$

The model optionally includes regularization on $R^{-1}$ and/or $C^{-1}$. An EM algorithm fits maximum penalized likelihood parameter estimates and, as a by-product, imputes missing entries of $Y$.

TRCMA, a one-step approximation to this EM algorithm, was proposed as a computationally tractable alternative. But even this approximation is much slower than all other methods we have worked with in this paper, especially for large $N$–all other methods that explicitly model sample relatedness are given $K$ and so can leverage a one-off eigendecomposition of $K$ to derive iterations that are linear or quadratic in $N$; in contrast, TRCMA has $O(N^3)$ iterations (though it presumably could be modified to use $K$, or just its eigenvectors, in a similar way). The computational expense is also partially due to the search over regularization parameters: for both precisions in the matrix normal, a penalty amount and type ($\ell_1$ or $\ell_2$) must be chosen.

We use two shortcuts to mitigate this computational expense. First, we use only $\ell_2$ penalization: it is much faster than $\ell_1$ (as conditional updates have analytic solutions instead of calls to `glasso`) and [1] found that the $\ell_2$ penalty worked well even when the true precision matrices were sparse. Second, we performed preliminary simulations to find a set of reasonable regularization parameters for the model to choose from via cross-validation. Specifically, we searched over $(\rho_{\text{row}}, \rho_{\text{column}}) \in$

$\mathcal{G} := 10^{\{-5,-3.5,-2,-0.5,1\}} \times 10^{\{-6,-4.5,-3,-1.5,0\}}$ in all our analyses. We regularly observed that TRCMA chose regularization parameters in the interior of this grid, suggesting that these ranges are, very roughly speaking, sufficiently wide.

While these two speedups will certainly attenuate accuracy–we could have tried $\ell_1$ regularization, tuned the range of $\mathcal{G}$ to each dataset and increased the density of $\mathcal{G}$–we hope our compromise between run time and accuracy is reasonable and representative of the typical choices of end users.

## 2.4 KNN: $k$-nearest neighbors

We use the function `impute.knn` from the R package `impute` as a non-paramteric imputation benchmark [26, 6]. We use the default parameters–including, in particular, $k = 10$–except we allow phenotypes with arbitrary amounts of missingness (by default, the program returns an error when phenotypes have $> 80\%$ missingness). The method finds the $k$-nearest neighbors for each phenotype and then imputes missing values to the average of their observed neighbors.

## 2.5 mice: multiple imputation by chained equations

We implement this method with the R package `mice` [27]. We use default parameters and average over 5 (the default value) multiply-imputed datasets; we have observed this performs dramatically better than simply taking the first imputed dataset.

`mice` implements a variety of imputation methods, but we only used predictive mean matching (`pmm`), the default for numeric variables. Iterating over phenotypes, the method predicts values for observed and missing samples using the other phenotypes and then matches each missing entry with the closest observed entries based on these predictions (we used the 5 closest matches, which is the default). Missing entries are then imputed to the observed value a randomly chosen partner.

The predictions on which matching is based are made by combining frequentist and Bayesian linear regression on covariates, $X$. In our implementation of the package, each phenotype $p$ is reqressed on all other phenotypes, so $X = \hat{Y}_{-p}$, where $\hat{Y}_{,-p}$ is the current imputed data matrix after removing phenotype $p$.

For observed entries, predictions are the OLS fitted values:

$$\hat{Y}_{obs,p} := X_{obs,}\hat{\beta}$$

where $\hat{\beta}$ is the MLE. The missing entries are also of the form $X\beta$, except now the regression coefficients $\beta^*$ are now drawn randomly from their posterior (using the default $\mathcal{N}\left(0, 10^{-5}I\right)$ prior):

$$\hat{Y}_{miss,p} := X_{miss,}\beta^*$$

## 2.6 softImpute

We use the softImpute method of [16] as a benchmark from the matrix completion literature in machine learning. We consider this method roughly representative of the state-of-the-art in this field [28, 15], though reported comparisons suggest that the relative performances of the many matrix completion methods depend heavily on the dataset.

softImpute maximizes the penalized likelihood

$$\min_M \sum_{n,p \in obs} (Y_{np} - M_{np})^2 + \lambda ||M||_*$$

where $||M||_*$ is the nuclear norm of $M$, or the $\ell_1$ norm of $M$'s singular values, and measures the complexity of $M$ and thus discourages overfitting. Since the $\ell_1$ penalty induces sparsity, the fitted $M$ typically has low rank, which is the key to softImpute's computational efficiency.

Our implementation follows the guide at

`http://web.stanford.edu/~hastie/swData/softImpute/vignette.html`

Specifically: we use the alternating least squares algorithm; we start with the maximum rank set to zero and then, as we shrink the regularization, allow the solution's rank to grow by at most two at each new $\lambda$; we vary $\log \lambda$ along 100 evenly spaced points on the interval $[-3 \log 10, \log(\lambda_0 + .2)]$, where $\lambda_0$ is the minimum $\lambda$ such that the solution, $\hat{M}_\lambda$, is 0; and we choose $\lambda$ by 10-fold cross validation to maximize predictive accuracy.

## 2.7 MPMM: multiphenotype mixed models

We fit MPMM by estimating the $B$ and $E$ parameters of model (2) on the rows of $Y$ that have been fully observed (i.e. case-wise deletion). We use our `R` implementation from [3], though the command line tool from [33] fits the same model in essentially the same way (modulo a Newton step once the EM algorithm has nearly converged).

Given observed phenotypes and variance component estimates, MPMM imputes missing entries to their conditional expectations, or BLUPs. Defining $\Sigma := (B \otimes K + E \otimes I_N)$,

$$\begin{aligned}
\mathbb{E}\left(y_{miss} | y_{obs}, B, E\right) &= \mathrm{Cov}\left(y_{miss}, y_{obs} | B, E\right) \mathbb{V}\left(y_{obs} | B, E\right)^{-1} y_{obs} \\
&= \Sigma_{miss,obs} \left[\Sigma_{obs,obs}\right]^{-1} y_{obs}
\end{aligned}$$

In general, these computations cost $O(|obs|^3)$ (or $O(|miss|^3)$ if a Schur complement identity is used), and thus the cost of imputing is $O(N^3 P^3)$ if some fixed fraction of entries are missing as $N$ and $P$ vary.

In the special case where samples are either entirely observed or entirely missing, the above conditional expectation can be computed in $O(N^3 + P^3)$. This is because, in this special case, the subsetting operations that select missing or observed entries commute with the Kronecker product structure. Specifically, if $M$ are missing samples and $O$ are observed samples, we can write, by assumption on the missingness pattern, $\mathrm{vec}\left(Y_{O,}\right) = y_{obs}$ and $\mathrm{vec}\left(Y_{M,}\right) = y_{miss}$, and so

$$\begin{aligned}
\mathbb{E}\left(y_{miss} | y_{obs}, C, D\right) &= \left(B \otimes K\right)_{miss,obs} \left[\left(B \otimes K + E \otimes I\right)_{obs,obs}\right]^{-1} y_{obs} \\
&= \left(B \otimes K_{MO}\right) \left[B \otimes K_{OO} + E \otimes I_{|O|}\right]^{-1} \mathrm{vec}\left(Y_{O,}\right) \\
&= \left(B^{1/2} \otimes K_{MO}\right) \left(\left[B^{-1/2} E B^{-1/2}\right] \oplus K_{OO}\right)^{-1} \mathrm{vec}\left(Y_{O,} B^{-1/2}\right)
\end{aligned}$$

By lemma 2, this can be computed in $O(N_O^2 P + N_O N_M P + P^3)$ (by retaining the eigendecomposition of $K_{OO}$ from the parameter learning step).

While this pattern of missingness will essentially never occur in a real dataset–and if it did one would prefer to drop unphenotyped samples since this results in no loss of phenotype data–it does occur in out-of-sample prediction problems, as discussed in [20].

# 3 Simulation descriptions

## 3.1 Simulations to assess phenotype imputation accuracy

The results presented in Figure 1 use data simulated from a standard MPMM. Defining `cov2cor` to map covariance matrices to their respective correlation matrices, we draw

$$Y = U + \epsilon \tag{21}$$

$$U \sim \mathcal{MN}\left(0, K, h^2 \texttt{cov2cor}(B)\right) \tag{22}$$

$$\epsilon \sim \mathcal{MN}\left(0, I, (1-h^2)\texttt{cov2cor}(E)\right) \tag{23}$$

We generally take $N = 300$, $P = 15$, $B$ to be an AR(1) matrix with autocorrelation $\rho = .45$ and $E \sim \text{Wi}\left(P, \frac{1}{P}I\right)$, with $E$ being redrawn for each simulated dataset. We use two types of $K$ matrices: either a block diagonal matrix with blocks corresponding to independent sets of 4 siblings or a random subsample, redrawn for each simulated datset, of the kinship matrix derived from the human NSPHS study [11]. Finally, 5% of entries are hidden, completely at random, and their values retained to assess imputation accuracy.

We refer to this as our baseline simulation, and Figure 1 shows the resulting imputation correlations for each method. Supplementary Figures 2-8 all take the same basic form, with each modifying one aspect of the baseline simulation and then plotting the resulting imputation accuracy as in Figure 1. The changes are explained in the plot captions or, when necessary, in the below text. For reference, the results of the baseline simulation from Figure 1 are plotted as dotted lines in the background.

We assessed $h^2$ at 11 evenly spaced points between .05 and .95. All methods were run on 250 independently simulated datasets for each value of $h^2$, and averages over these 250 replicates are plotted in all figures. Two hours on a server was more than enough time for all methods to run the 2,750=$11 \times 250$ datasets, with two exceptions: TRCMA ran only $\approx 125$ datasets in the same amount of time and, for the larger data size in Supplementary Figure 3, we ran methods for four hours (LMM still only ran $\approx 1500$ datasets and TRCMA ran none).

## 3.2 Cancellation of genetic and environmental covariances

Simulation results shown in Figure 1 of the main paper suggest that performance generally decreases as heritability increases, but slightly increases at very high levels of heritability. Our hypothesis was that this occured due to cancellation of genetic and environmental covariances. To investigate this we repeated the simulations in Figure 1 with a different model for the genetic covariance ($B$ in (22)) with opposing genetic and environmental correlations i.e. $B_{pq} = -E_{pq}$ for $p \neq q$. In this model, the cancellation is exact at $h^2 = .5$, in that $\mathbb{V}(Y_{i,})$ is diagonal for all $i$. The results are shown in Supplementary Figure 2. For moderate $h^2$, genetic and environmental correlations cancel, impeding imputation for multitrait methods relative to the dotted lines, which show the results from Figure 1. At large $h^2$, the cancellation effect is outweighed by the increased size of $|B_{pq}|$ and so imputation improves.

## 3.3 Effect of non-random missingness

Our model implicitly assumes that missingness is ignorable in the update for $Q_Y$ (equations (8) and (9)) and we simulate this in our baseline by removing 5% of entries uniformly at random. We can

simulate data with non-ignorable missingness, however, by removing entries of $Y$, independently, with probability depending on the values of the entries:

$$P(\text{ entry } (i,j) \text{ is missing }) \propto \Phi(Y_{ij})$$

where $\Phi$ is the standard normal cdf. The proportionality constant is chosen to ensure 5% overall missingness (in expectation over the random missingness pattern).

### 3.4 Effect of unmodelled shared environment

We investigated the performance of the different methods in the presence of (unmodelled) shared environmental effects. To do this we added a random effect representing shared environment to the simulated data, in addition to the genetic relatedness and idiosyncratic noise random effects in a standard MPMM:

$$\begin{aligned} Y &= a^2U + c^2C + e^2\epsilon \\ U &\sim \mathcal{MN}\left(0, K, \texttt{cov2cor}(B)\right) \\ C &\sim \mathcal{MN}\left(0, R, \texttt{cov2cor}(D)\right) \\ \epsilon &\sim \mathcal{MN}\left(0, I, \texttt{cov2cor}(E)\right) \end{aligned}$$

Such models are often called ACE models, where $U$ is the Additive efffect, $C$ is a Common environmental effect and $\epsilon$ is the purely independent Environmental contribution [4].

We take $K$, $B$ and $E$ as in the baseline model and $D$ is drawn (independently) from the same distribution as $E$ for each simulated dataset. We define $R$ to be block diagonal with 10 independent environments and each block/environment to be an AR(1) matrix with autocorrelation $\rho = .5$.

Defining the heritability as $h^2 = (a^2 + c^2)/(a^2 + c^2 + e^2)$ and fixing the relative sizes of $a^2$ and $c^2$ to three different values given in the caption, the x-axis in Supplementary Figure **??** determines the relative contributions of the unstructured $\epsilon$ and the structured $U$ and $C$.

### 3.5 Effect of non-normally distributed phenotypes

To create non-normal phenotypes, we start with the baseline MPMM but transform the noise:

$$Y = U + \left(\exp(\epsilon_{ij})\right)_{ij}$$

Phenotype imputation is then performed either on $Y$ or on a quantile normalized version; quantile normalization is natural for most downstream analyses, including GWAS.

### 3.6 Type I error calibration

To assess the impact of phenotype imputation on the null distribution of p-values in a GWAS, we simulated phenotype data from an MPMM with no genetic contribution beyond the background term $U$. We imputed missing data and then tested the resulting phenotypes against SNP data and assessed the null distribution of the resulting p-values (Supplementary Figure 9).

We present results for simulations with $N = 300$, $P = 15$, $h^2 = .2$, $B$ an AR(1) with autocorrelation parameter $\rho = .2$ and $E \sim \text{Wi}\left(P, \frac{1}{P}I\right)$; we note the results did not qualitatively change when varying $\rho \in \{-.2, .2, .5\}$ and $h^2 \in \{.1, .2, .5\}$. We chose two types of $K$ matrix, one corresponding

to independent sets of 4 siblings and one a random subsample of the kinship matrix derived from the human NSPHS study [11]. We then added 10% missingness and either dropped missing samples in testing (Unimputed) or imputed with PHENIX, MVN or MPMM; we note the results did not qualitatively change for missingness levels in $\{.01, .05, .1, .2, .5\}$.

We tested both real and simulated genotypes. For the sibling $K$ simulations, we generated SNPs in a hierarchical way: first, we drew parental alleles independently and then we simulated sibling genotypes via Mendel's rules. We simulated 100,000 unlinked loci on which we performed GWAS, for each of the $P = 15$ phenotypes, with `gemma` using the default QC filters (top row of Supplementary Figure 9) [32].

For the simulations where $K$ is a subset of the NSPHS dataset, we used real SNPs corresponding to the same subset of the NSPHS dataset. SNPs were imputed (see Online Methods) and we performed GWAS on the resulting 9,165,236 SNPs with `gemma` using the default QC filters (bottom row of Supplementary Figure9) for each of the 15 phenotypes.

### 3.7 Power of single phenotype tests

We performed a simulation study to assess the power gains from phenotype imputation. We simulated data using a standard MPMM as before, except now we add a causal SNP:

$$\begin{aligned} Y &= X\beta + U + \epsilon \\ U &\sim \mathcal{MN}(0, K, B) \\ \epsilon &\sim \mathcal{MN}(0, I, E) \end{aligned}$$

We choose $N = 5,000$ and $P = 15$. We also choose $B$ to be AR(1) with autocorrelation parameter $\rho = -.2$ so that, in particular, there is a mixture of positive and negative genetic correlations amongst the phenotypes. We again take $E \sim \text{Wi}\left(P, \frac{1}{P}I\right)$ except now we do not resample $E$ for each dataset but rather fix it at the outset (though $U$ and $\epsilon$ are still randomly drawn for each dataset). We choose $K$ to represent independent sets of 4 siblings. $X \in \mathbb{R}^N$ is a common SNP that we draw independently for each dataset by $X_i \overset{\text{iid}}{\sim} \text{Binomial}(2, .2)$.

We choose a pleiotropic $\beta$ so that the SNP $X$ has a substantial effect on the first phenotype, which represents a phenotype of primary interest, and lesser but non-negligible effects on the other fourteen phenotypes, which represent phenotypes related to and collected with the first, primary phenotype. In this section, we are interested only in the first phenotype, and the other fourteen are valuable only as a means for imputing missing entries in the first. Specifically, we choose $\beta$ in terms of the implied percent variance explained (PVE) in each of the phenotypes: the PVE for phenotype 1 is 8%, and the other 14 PVEs were drawn randomly:

$$\text{PVE}_{2:15} \overset{\text{iid}}{\sim} \frac{2\text{PVE}_1}{3}\left|\mathcal{N}(0,1)\right|$$

To introduce sparsity into $\beta$, the smallest 5 PVE values were then hard-thresholded to 0. The realized values used to create Supplementary Figure 10 are displayed in the first columns of the below table.

| | Univariate Test | | MV Test, One | | MV Test, Sparse | | MV Test, Dense | |
|---|---|---|---|---|---|---|---|---|
| Phenotype | PVE | Coeff | PVE | Coeff | PVE | Coeff | PVE | Coeff |
| 1 | 8.00 | 0.28 | 8.00 | 0.28 | 7.30 | 0.27 | 6.00 | 0.24 |
| 2 | 2.70 | 0.16 | 0.00 | 0.00 | 2.40 | 0.15 | 2.00 | 0.14 |
| 3 | 2.30 | 0.15 | 0.00 | 0.00 | 2.10 | 0.14 | 1.70 | 0.13 |
| 4 | 5.40 | 0.23 | 0.00 | 0.00 | 4.90 | 0.22 | 4.00 | 0.20 |
| 5 | 1.90 | -0.14 | 0.00 | 0.00 | 1.70 | -0.13 | 1.40 | -0.12 |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 7 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.30 | -0.05 |
| 8 | 7.60 | -0.28 | 0.00 | 0.00 | 7.10 | -0.27 | 5.90 | -0.24 |
| 9 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.03 |
| 10 | 3.40 | 0.18 | 0.00 | 0.00 | 3.10 | 0.18 | 2.60 | 0.16 |
| 11 | 5.90 | -0.24 | 0.00 | 0.00 | 5.50 | -0.23 | 4.60 | -0.21 |
| 12 | 2.10 | -0.14 | 0.00 | 0.00 | 1.90 | -0.14 | 1.60 | -0.13 |
| 13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.70 | -0.08 |
| 14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.60 | 0.08 |
| 15 | 4.90 | -0.22 | 0.00 | 0.00 | 4.50 | -0.21 | 3.80 | -0.19 |

Randomly generated PVEs and corresponding regression coefficients used to generate Supplementary Figures 10 (first 2 columns) and 11 (last six columns, each pair corresponding to a different line type in S11) for 15 simulated phenotypes. Univariate tests (columns 1 and 2) are performed on phenotype 1; mulitvariate tests (rows 3-8) are performed on all 15 phenotypes. The first entry in each column is non-random while all others were drawn randomly (once) and fixed to the resulting values for all simulated datasets.

### 3.8 Power of multiple phenotype tests

For each SNP of interest at a time, we use a multi-phenotype mixed model (MPMM) to test association with a set of $P$ phenotypes:

$$\begin{aligned} Y &= X\beta + U + \epsilon \\ U &\sim \mathcal{MN}(0, K, B) \\ \epsilon &\sim \mathcal{MN}(0, I, E) \end{aligned}$$

where $X \in \mathbb{R}^{N\times 1}$ is the vector of genotypes. Specifically, we test $\beta = 0_P$ with the likelihood ratio

$$\text{LRT} = -2\left(ll(\beta = 0, \hat{B}_0, \hat{E}_0) - ll(\beta = \hat{\beta}, \hat{B}_1, \hat{E}_1)\right)$$

where $ll$ is the log-likelihood in the above MPMM and all estimated parameters are MLEs.

Forming the LRT requires fitting variance components ($B$'s and $E$'s), estimating $\beta$ and evaluating log-likelihoods. Due to the cost of fitting the variance components, we fit fit only $\hat{B}_0$ and $\hat{E}_0$ and then make the approximation $(\hat{B}_0, \hat{E}_0) = (\hat{B}_1, \hat{E}_1)$. Because

$$\max_{\beta,B,E} ll(\beta, B, E) \geq \max_{\beta} ll(\beta, \hat{B}_0, \hat{E}_0) = ll\left(\hat{\beta}\left(\hat{B}_0, \hat{E}_0\right), \hat{B}_0, \hat{E}_0\right)$$

the approximate LRT lower-bounds the exact LRT and our method is conservative. Nonetheless, this approximation is expected to be good for typical analyses, where individual SNPs are expected

to explain a nearly negligible fraction of the overall variance; however, it may attenuate power when analyzing SNPs with very large effect sizes [32].

### 3.8.1 Simulation details

As in the univariate simulations for Supplementary Figure 10, we choose $N = 5,000$, $P = 15$, $B$ to be AR(1) with autocorrelation parameter $\rho = -.2$, $K$ to represent independent sets of 4 siblings and we draw the common SNP, independently for each dataset, by $X_i \overset{\text{iid}}{\sim} \text{Binomial}\,(2, .2)$. We also take the same $E$ from the univariate simulations, which was drawn $\text{Wi}\left(P, \frac{1}{P}I\right)$

We use three different choices for $\beta$ in this section to represent varying levels of pleiotropy. In the first situation (UV signal), the causal SNP affects only the first phenotype; in the second (sparse), the SNP affects some (10), but not all, of the phenotypes; in the third (dense), the SNP affects all (15) phenotypes. All 15 phenotypes are tested for association with the SNP $X$.

We again parameterize our choices for $\beta$ in terms of the implied PVE. For the first simulation set the PVE to 8% for the first phenotype (and 0 for the others). The other PVEs were derived from the univariate test power simulations: the dense and sparse PVEs were proportional to the PVEs drawn in the previous section prior to and after, respectively, the hard-thresholding step. Proportionalty constants were chosen to yield power away from 0 and 1 (for the tests without added missingness). The resulting PVEs and effect sizes are displayed in the table in Section 3.7.

### 3.8.2 Computational simplification

In general, the normal equation for regressing the response $y$ on covariates $X$ with noise precision $\Omega$ is

$$\hat{\beta}^{MLE} = \left(X^T \Omega X\right)^{-1} X^T \Omega y$$

In our application, we take the covariates to be $I_P \otimes X \in \mathbb{R}^{NP \times P}$ ($X \in \mathbb{R}^{N \times 1}$ by assumption), the response to be $\text{vec}\,(Y) \in \mathbb{R}^{NP}$, and the noise precision, which incorporates the heritable random effect, to be

$$\Omega = (B \otimes K + E \otimes I_N)^{-1} = (L \otimes Q)\,\Lambda^{-1}\,(L \otimes Q)^T \tag{24}$$

where $Q\Lambda_N Q^T$ is an eigendecomposition of $K$; $Q_P \Lambda_P Q_P^T$ is an eigendecomposition of $B^{-1/2} E B^{-1/2}$; $L := B^{-1/2} Q_P$; $\Lambda := \Lambda_P \oplus \Lambda_N$. This decomposition is closely related to those in [5, 33, 20].

Returning to the normal equation and plugging in the MPMM-specific values for $y$, $X$ and $\Omega$,

$$\begin{aligned}
\hat{\beta}^{MLE} &= \left((I_P \otimes X)^T \left[(L \otimes Q)\,\Lambda^{-1}\,(L \otimes Q)^T\right](I_P \otimes X)\right)^{-1} (I_P \otimes X)^T \left[(L \otimes Q)\,\Lambda^{-1}\,(L \otimes Q)^T\right] y \\
&= L^{-T} \underbrace{\left(\left(I \otimes [Q^T X]^T\right)\Lambda^{-1}\left(I \otimes [Q^T X]\right)\right)^{-1}}_{\Omega_X} \left(I \otimes \underbrace{[Q^T X]}_{X'}{}^T\right) \text{vec}\left(\underbrace{\left[\text{mat}(\Lambda^{-1}) * \left(Q^T Y L\right)\right]}_{Z}\right) \\
&= L^{-T} \Omega_X \text{vec}\left(X'^T Z\right) \\
&= X'^T Z \Omega_X L^{-1}
\end{aligned}$$

Because we only test one covariate at a time, $\Omega_X$ is just a $P \times P$ matrix (if, instead, $D > 1$ covariates are used, this becomes a $DP \times DP$ matrix and requires partial trace operations). In

fact, $\Omega_X$ is diagonal with

$$\left((\Omega_X)_{pp}\right)^{-1} = X^T Q \left[\Lambda^{-1}\right]_{(pp)} Q^T X = || \left[\Lambda^{-1/2}\right]_{(pp)} X' ||_2^2$$

which is manageable since $\Lambda$ is diagonal.

Once $\hat{\beta}$ is evaluated, the likelihood can be compactly evaluated for both $Y$ and $Y - X\hat{\beta}$ using previous results [3].

### 3.9 Calibrating the imputation metric $r$

To assess the calibration of our imputation metric $r$, we simulated from our baseline model and compared the true and estimated imputation correlations. We averaged over 1,000 independently simulated datasets. The results are shown in Supplementary Figure 12. The black lines in the top row show the true imputation correlation using our oracle knowledge of the heldout, simulated data, and are essentially identical to the red lines in Figure 1 (we only consider PHENIX in these assessments).

The brown and purple lines show two different estimators for $r$, which in practice is unknown since the missing data is truly unobserved. Both estimators are formed by first hiding some of the entries of $Y^o$, the observed part of $Y$, to form $\tilde{Y}^o$. This new phenotype matrix is then imputed, returning a fully-observed matrix $\hat{Y}$. Finally, $r$ is estimated as the correlation between $\hat{Y}$ and $Y^o$ at the entries hidden from $Y^o$ to create $\tilde{Y}^o$.

The brown and purple lines differ by $f$, the fraction of $Y^o$ masked to create $\tilde{Y}^o$. As $f \to 1$, $\tilde{Y}^o$ becomes a completely blank matrix and phenotype imputation becomes impossible, yielding estimates of $r$ near 0; conversely, as $f \to 0$, a vanishingly small number of entries of $Y^o$ are masked, resulting in highly variable estimates of $r$.

We have plotted two choices for $f$ that compromise between this bias at $f = 1$ and variance at $f = 0$. The additional bias from choosing the larger $f$ explains the gap between the purple and brown lines in the top row of Supplementary Figure 12, though even the brown lines are slightly downwardly biased. The additional variance coming from the smaller choice of $f$ is evident but mitigated by our averaging over many simulated datasets. Ultimately, despite this bias and variance, the bottom row of Supplementary Figure 12 shows that our estimates of $r$ are very close and, at worst, conservative.

In practice it is possible to average these $r$ estimates across many replicates of the masking process to create $\tilde{Y}^o$ from $Y^o$, leading to estimates with lower variance (and thus making choices of small $f$ feasible). In our GWAS, for example, we repeated this sub-sampling 10,000 times with $f = .05$ to remove essentially all sub-sampling variance.

Though this procedure is involved, it is easy to implement in our `R` package. Moreover, this procedure can be performed phenotype-wise, computing imputation correlations within-phenotype and returning a vector of $r$'s. This vector can be used to inform downstream analyses, as we did in our rat GWAS analysis and can be seen in Figure 3.

### 3.10 Runtimes on simulated and real datasets

Most (method, dataset) pairs were run on 64 2.30 GHz processors (AMD Opteron 6276) in parallel for 12 hours or until all 3,000 simulated missingness patterns had run (100 for each of 30 levels of

added missingness). We made exceptions for the particularly computationally expensive (method, dataset) pairs.

First, MPMM and TRCMA were dramatically more costly than other methods, and so were only run on NSPHS and wheat, two of the smaller datasets (on 64 2.30 GHz processors (AMD Opteron 6276) and 16 3.30GHz processors (Intel Xeon E5-2667) in parallel, respectively). For both these datasets, we ran MPMM on all 3,000 simulated missingness patterns (though it's case-wise deletion approach discarded all data and could not run for 75% and 50% of the patterns in NSPHS and wheat, respectively).

Next, for (TRCMA, NSPHS), by far the most expensive situation studied, we ran on five missingness patterns for each level of missingness below 20%; above this cutoff, one missingness pattern was run for each missingness level. For (TRCMA,wheat) we ran 6 or 7 missingness patterns for each missingness level.

Finally, the chicken dataset had far greater $N$ than any other dataset, which caused LMM and PHENIX–the methods using relatedness–to become far more expensive; for example, a full-rank eigendecomposition of $K$ costs roughly a half hour. We run both these methods on 16 3.30GHz processors (Intel Xeon E5-2667) in parallel for 20 independent missingness patterns at 15 missingness levels (giving 300, rather than 3,000, simulated datasets) without any time constraints.

We note that we could have pre-computed the eigendecomposition of $K$ for PHENIX but not for LMM; the former does not drop samples and thus always works with the same $K$ while the latter drops a different set of samples for each phenotype and thus performs $P$ unique eigendecompositions. For sufficiently large $N$, this means that performing $P$ LMMs will be $P$ times more expensive than PHENIX, meaning our new method would be both more powerful and much faster.

| | $N$ | $P$ | phenix | MVN | LMM | softI | KNN | mice | MPMM | TRCMA |
|---|---|---|---|---|---|---|---|---|---|---|
| UK BS | 1,500 | 6 | 0.8 | 0.1 | 0.9 | 0.3 | 0 | 0.1 | | |
| NSPHS | 1,021 | 15 | 1.2 | 0.1 | 1 | 0.4 | 0 | 0.1 | 100.8 | 144 (h) |
| Wheat | 720 | 7 | 0.2 | 0 | 0.1 | 0.2 | 0 | 0 | 0.5 | 8 (h) |
| Rats | 1,407 | 140 | 131.2 | 3.5 | 16.3 | 22.9 | 0 | 9.7 | | |
| Yeast | 1,008 | 46 | 5.1 | 0.2 | 2.6 | 2.4 | 0 | 0.7 | | |
| Chickens | 11,575 | 14 | 89.5 | 0.8 | 154.2 | 4.2 | 0 | 4 | | |
| Fig 1 | 300 | 15 | 0.1 | 0 | 0.1 | 0.1 | 0 | 0.1 | 7 | 41 |
| Fig S3 | 1,000 | 50 | 3.9 | 0.1 | 9.3 | 2.2 | 0 | 0.9 | | |

Average runtimes for each method on each dataset. Times are in minutes by default, but (h) means the time is in hours. Except TRCMA, MPMM and, on the chicken dataset only, phenix and LMM, all running times were recorded in identical computing environments.

# 4 Appendix: Jeffreys' prior for matrix factorization

We use a matrix factorization model as our prior on the genetic contribution $U$:

$$U = S\beta;\ S \sim \mathcal{MN}(0, K, I);\ \beta \sim \mathcal{MN}\left(0, I, \tau^{-1} I\right)$$

As $\tau \to 0$, the prior on $\beta$ becomes flat (also called objective, or non-informative, because such priors typically deliver unregularized estimates). In contrast, as $\tau \to 0$, the implied prior on $U$

*does* become flatter, but *does not* become flat. This means that even in the improper limit of $\tau = 0$–which we use as a default–our prior still encourages $U$ to shrink toward the prior mean of 0.

[17] shows this using the invariance property of Jeffreys priors. First, the Jeffreys prior on $U$ is flat, and therefore the Jeffreys prior on $(S, \beta)$ induces a flat prior on $S\beta$. But the (improper) Jeffreys prior on $(S, \beta)$ is, when $N = M = P = 1$,

$$p(S, \beta) \propto \sqrt{S^2 + \beta^2}$$

As $\tau \to 0$, the concave, normal priors that we uses to model $S$ and $\beta$ become flatter and thus closer to this strictly convex, quadratic Jeffreys prior. This explains why choosing small $\tau$ minimizes shrinkage, but it also explains why even $\tau = 0$ cannot eliminate shrinkage.

We derive the Jeffreys prior for general $N$, $M$ and $P$ below.

**Proposition 1.** *Let*

$$Y \sim \mathcal{MN}(S\beta, I, I) \tag{25}$$

*Then the prior on $(S, \beta)$ which induces a flat prior on $S\beta$ is*

$$p(S, \beta) \propto \sqrt{|S^T S|^{P-M} |\beta\beta^T|^{N-M} |(S^T S) \oplus (\beta\beta^T)|}$$

*Proof.* Following [17], we first show that the flat prior on $U$ is the Jeffreys prior on $U$; then, since the Jeffreys prior is invariant under reparameterization, the Jeffreys prior on $U$ is equivalent to the Jeffreys prior on $(S, \beta)$. This shows the Jeffreys prior on $(S, \beta)$ induces a flat prior on $U$.

First, reparameterize the likelihood in terms of $U := S\beta$, so that

$$\ell(Y|U) \equiv -\frac{1}{2}||(Y - U)||_F^2$$

Since this log likelihood is quadratic, the Hessian with respect to $U$ is constant, thus so is its expectation, the Fisher information. Because the Jeffreys prior on $U$ depends only on the Fisher information, it, too, must be constant. Then, since the Jeffreys prior on $(S, \beta)$ necessarily induces the Jeffreys prior on $U$, the Jeffreys prior on $(S, \beta)$ induces the flat prior on $U$.

Finding the Fisher information requires the log-likelihood derivatives:

$$\begin{aligned}
\frac{\partial \ell(Y|S, \beta)}{\partial S} &= -Y\beta^T + S\beta\beta^T \\
\frac{\partial \ell(Y|S, \beta)}{\partial \beta} &= -S^T Y + S^T S\beta
\end{aligned}$$

This leads to expected second derivatives

$$\begin{aligned}
\frac{\partial}{\partial S_{im}} \frac{\partial \ell(Y|S, \beta)}{\partial S} &= I_{im}\beta\beta^T \implies \nabla_S^2 \ell(Y|S, \beta) = (\beta\beta^T) \otimes I_N \\
\frac{\partial}{\partial \beta_{mp}} \frac{\partial \ell(Y|S, \beta)}{\partial \beta} &= S^T S I_{mp} \implies \nabla_\beta^2 \ell(Y|S, \beta) = I_P \otimes (S^T S) \\
\frac{\partial}{\partial \beta_{mp}} \frac{\partial \ell(Y|S, \beta)}{\partial S} &= -Y I_{mp}^T + S\beta I_{mp}^T + S I_{mp}\beta^T \\
\implies \mathbb{E}\left( \frac{\partial}{\partial \beta_{mp}} \frac{\partial \ell(Y|S, \beta)}{\partial S} | S, \beta \right) &= S I_{mp}\beta^T \implies \mathbb{E}\left(\nabla_S \nabla_\beta \ell(Y|S, \beta)\right) = \beta \otimes S
\end{aligned}$$

and so the Fisher information is

$$\mathcal{I}\left(\mathrm{vec}\left(S\right),\mathrm{vec}\left(\beta\right)\right)=\left(\begin{array}{cc}\left(\beta\beta^{T}\right)\otimes I_{N} & \beta\otimes S\\ \beta^{T}\otimes S^{T} & I_{P}\otimes\left(S^{T}S\right)\end{array}\right) \tag{26}$$

Now the goal is to find the eigenvalues of $\mathcal{I}$. Let $\beta=U_{\beta}D_{\beta}V_{\beta}^{T}$ and $S=U_{S}D_{S}V_{S}^{T}$ be SVDs and whiten $\mathcal{I}$ by conjugating with the orthogonal matrix $U:=\left(U_{\beta}\otimes U_{S}\right)\times\left(V_{\beta}\otimes V_{S}\right)$, where $\times$ is the Cartesian product (or direct sum; we use non-standard notation because we reserve $\oplus$ for the Kronecker sum in this paper):

$$U^{T}\mathcal{I}U=\left(\begin{array}{cc}D_{\beta}D_{\beta}^{T}\otimes I_{N} & D_{\beta}\otimes D_{S}\\ D_{\beta}^{T}\otimes D_{S}^{T} & I_{P}\otimes D_{S}^{T}D_{S}\end{array}\right)=:\mathcal{I}'$$

Define $\Lambda_{\beta}=D_{\beta}D_{\beta}^{T}$ and $\Lambda_{S}=D_{S}^{T}D_{S}$ and let $\lambda_{i}^{S}=\left(\Lambda_{S}\right)_{ii}$, $\lambda_{i}^{\beta}=\left(\Lambda_{\beta}\right)_{ii}$ . Then the eigenvalues of $\mathcal{I}$ are roots of the characteristic polynomial:

$$\begin{aligned}
&\left|\mathcal{I}-\lambda I_{M^{2}NP}\right|=\left|\mathcal{I}'-\lambda I_{M^{2}NP}\right|\\
&=\left|\left(\begin{array}{cc}\left(\Lambda_{\beta}-\lambda I_{M}\right)\otimes I_{N} & D_{\beta}\otimes D_{S}\\ D_{\beta}^{T}\otimes D_{S}^{T} & I_{P}\otimes\left(\Lambda_{S}-\lambda I_{M}\right)\end{array}\right)\right|\\
&=\left|\left(\Lambda_{\beta}-\lambda I_{M}\right)\otimes I_{N}\right|\left|I_{P}\otimes\left(\Lambda_{S}-\lambda I_{M}\right)-\left(D_{\beta}^{T}\otimes D_{S}^{T}\right)\left(\left(\Lambda_{\beta}-\lambda I_{M}\right)\otimes I_{N}\right)^{-1}\left(D_{\beta}\otimes D_{S}\right)\right|\\
&=\left(\prod_{m}(\lambda_{m}^{\beta}-\lambda)\right)^{N}\left|I_{P}\otimes\left(\Lambda_{S}-\lambda I_{M}\right)-\left(\left[\Lambda_{\beta}\left(\Lambda_{\beta}-\lambda I\right)^{-1}\right]\times 0_{P-M,P-M}\right)\otimes\Lambda_{S}\right|\\
&=\left(\prod_{m}(\lambda_{m}^{\beta}-\lambda)\right)^{N}\prod_{m=1}^{M}\prod_{p=1}^{P}\left[\left(\lambda_{m}^{S}-\lambda\right)-I\{p\leq M\}\left(\frac{\lambda_{p}^{\beta}}{\lambda_{p}^{\beta}-\lambda}\right)\lambda_{m}^{S}\right]\\
&=\left(\prod_{m}(\lambda_{m}^{\beta}-\lambda)\right)^{N}\left(\prod_{m}\left(\lambda_{m}^{S}-\lambda\right)\right)^{P-M}\prod_{m,m'=1}^{M}\left[\left(\lambda_{m}^{S}-\lambda\right)-\lambda_{m}^{S}\left(\frac{\lambda_{m'}^{\beta}}{\lambda_{m'}^{\beta}-\lambda}\right)\right]\\
&=\left(\prod_{m}(\lambda_{m}^{\beta}-\lambda)\right)^{N-M}\left(\prod_{m}\left(\lambda_{m}^{S}-\lambda\right)\right)^{P-M}\prod_{m,m'=1}^{M}\left[\left(\lambda_{m}^{S}-\lambda\right)\left(\lambda_{m'}^{\beta}-\lambda\right)-\lambda_{m}^{S}\lambda_{m'}^{\beta}\right]\\
&=\left(\prod_{m}(\lambda_{m}^{\beta}-\lambda)\right)^{N-M}\left(\prod_{m}\left(\lambda_{m}^{S}-\lambda\right)\right)^{P-M}\prod_{m,m'=1}^{M}\left[\left(\lambda-(\lambda_{m}^{S}+\lambda_{m'}^{\beta})\right)\lambda\right]\\
&=\left|\Lambda_{\beta}-\lambda I\right|^{N-M}\left|\Lambda_{S}-\lambda I\right|^{P-M}\left|\lambda I-\Lambda_{\beta}\oplus\Lambda_{S}\right|\lambda^{M^{2}}
\end{aligned}$$

As in Appendix 1 of [17], I take the Jeffreys prior proportional to the square root of the product of non-zero eigenvalues of the Fisher information.

□

# 5 Appendix: Useful Linear Algebra Idenitities

**Lemma 1.** *Let $A\in\mathbb{R}^{P\times P}$ and $X\in\mathbb{R}^{N\times N}$. Then $tr_{P}\left(A\oplus X\right)^{-1}$ can be computed in $O(NP+P^{3})$ given the matrix of eigenvalues of $X$, $\Lambda_{X}$.*

*Proof.* First,

$$\mathrm{tr}_P\left(A \oplus X\right)^{-1} = \mathrm{tr}_P\left[\left(U_A \otimes U_X\right)\left(\Lambda_A \oplus \Lambda_X\right)^{-1}\left(U_A \otimes U_X\right)^T\right] = U_A\left[\mathrm{tr}_P\left(\left(\Lambda_A \oplus \Lambda_X\right)^{-1}\right)\right]U_A^T$$

To compute the right hand side, the eigendecomposition of $A$ must be performed ($O(P^3)$), an $NP$ diagonal matrix must be inverted ($O(NP)$) and partial-traced out ($O(NP)$), and finally $P \times P$ matrix multiplications are performed ($O(P^3)$).

□

**Lemma 2.** *Let $A \in \mathbb{R}^{P\times P}$, $X \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times P}$ and let $X = Q_X\Lambda_X Q_X^T$ be a known eigendecomposition of $X$. Then $\left[A \oplus X\right]^{-1}$ vec$(B)$ can be computed in:*

- $O(P^3 + N^2P)$ *in general*
- $O(P^3 + NP^2)$ *if $X$ is diagonal*
- $O(P^3 + RP^2 + RNP)$ *if $X$ has rank $R$*
- $O(P^3 + RP^2)$ *if $X$ is diagonal and has rank $R$*

*Proof.* First,

$$\left(\left[A \oplus X\right]^{-1}\right)\mathrm{vec}\left(B\right) = \left(U_A \otimes Q_X\right)\underbrace{\left[\Lambda_A \oplus D\right]^{-1}\mathrm{vec}\left(Q_X^T B U_A\right)}_{\mathrm{vec}(Z)} = \mathrm{vec}\left(Q_X Z U_A^T\right)$$

There are four types of operations above

1. eigendecomposition of $A$
2. multiplication of an $N \times P$ matrix with a $P \times P$ matrix ($BU_A$ and $ZU_A^T$)
3. matrix multiplication an $N \times N$ matrix with an $N \times P$ ($Q_X^T B$ and $Q_X Z$)
4. diagonal $NP \times NP$ matrix operations

In general, 1 costs $O(P^3)$; 2 costs $O(NP^2)$; 3 costs $O(N^2P)$; and 4 costs $O(NP)$. When $X$ is diagonal, $Q_X = I$ and 3 can be elided. If $X$ is low-rank, $B$ and $Z$ can be compressed to $\mathbb{R}^{R\times P}$ and the cost of 2 becomes $O(RP^2)$; analogously, 3 becomes $O(RNP)$ and 4 becomes $O(RP)$. Finally, if additionally $X$ is diagonal, 3 can again be skipped.

□

# References


[1] Genevera I Allen and Robert J Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764–790, June 2010.

[2] Emamnuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[3] Andy Dahl, Victoria Hore, Valentina Iotchkova, and Jonathan Marchini. Network inference in matrix-variate Gaussian models with non-independent noise. *arXiv:1312.1622v1*, pages 1–17, December 2013.

[4] D.S. Falconer and Trudy Mackay. *Introduction to Quantitative Genetics*.

[5] Nicholas A Furlotte and Eleazar Eskin. Efficient multiple trait association and estimation of genetic correlation using the matrix-variate linear mixed-model. *Genetics*, pages genetics–114, 2015.

[6] Trevor Hastie, Robert Tibshirani, and Gavin Sherlock. Imputing missing data for gene expression arrays. *Technical Report, Division of Biostatistics, Stanford University*, pages 1–9, 1999.

[7] Alexander Ilin and Tapani Raiko. Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, 11:1957–2000, 2010.

[8] Hyun Min Kang, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–23, March 2008.

[9] Yong-deok Kim and Seungjin Choi. Variational Bayesian view of weighted trace norm regularization for matrix factorization. *IEEE Signal Processing Letters*, 20:261–264, 2013.

[10] Arthur Korte, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics*, 44(9):1066–1071, 2012.

[11] Gordan Lauc, Abdelkader Essafi, Jennifer E. Huffman, Caroline Hayward, Ana Knežević, Jayesh J. Kattla, Ozren Polašek, Olga Gornik, Veronique Vitart, Jodie L. Abrahams, Maja Pučić, Mislav Novokmet, Irma Redžić, Susan Campbell, Sarah H. Wild, Fran Borovečki, Wei Wang, Ivana Kolčić, Lina Zgaga, Ulf Gyllensten, James F. Wilson, Alan F. Wright, Nicholas D. Hastie, Harry Campbell, Pauline M. Rudd, and Igor Rudan. Genomics meets glycomics-the first gwas study of human N-glycome identifies HNF1A as a master regulator of plasma protein fucosylation. *PLoS Genetics*, 6(12):1–14, 2010.

[12] Yew Jin Lim and Yee Whye Teh. Variational Bayesian approach to movie rating prediction. *Proceedings of KDD Cup and Workshop*, 7:15–21, 2007.

[13] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M CM Kadie, Robert I Davidson, and David Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, January 2011.

[14] Roderick J A Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, NY, USA, 1986.

[15] Dehua Liu, Tengfei Zhou, Hui Qian, Congfu Xu, and Zhihua Zhang. A nearly unbiased matrix completion approach. In *Learning and Knowledge Discovery in Databases*, pages 210–225. Springer Berlin Heidelberg, 2013.

[16] Rahul Mazumder, Trevor Hastie, and Robert J Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, March 2010.

[17] Shinichi Nakajima and Masashi Sugiyama. Theoretical analysis of Bayesian matrix factorization. *The Journal of Machine Learning Research*, 12:2583–2648, 2011.

[18] Shinichi Nakajima, Masashi Sugiyama, S Derin Babacan, and Ryota Tomioka. Global analytic solution of fully-observed variational Bayesian matrix factorization. *The Journal of Machine Learning Research*, 14(1):1–37, 2013.

[19] Shinichi Nakajima, R Tomioka, M Sugiyama, and S Derin Babacan. Perfect Dimensionality Recovery by Variational Bayesian PCA. *Advances in Neural Information Processing Systems*, pages 971–979, 2012.

[20] Barbara Rakitsch, Christoph Lippert, Karsten Borgwardt, and Oliver Stegle. It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals. In *Advances in Neural Information Processing Systems*, pages 1466–1474, 2013.

[21] Benjamin Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

[22] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. *International Conference on Machine Learning*, pages 880–887, 2008.

[23] Vincent Segura, Bjarni J Vilhjálmsson, Alexander Platt, Arthur Korte, and Magnus Nordborg. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics*, 44(7):825–830, July 2012.

[24] Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS computational biology*, 6(5):e1000770, May 2010.

[25] Matthew Stephens and David J Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, October 2009.

[26] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, Trevor Hastie, Robert J Tibshirani, David Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–5, June 2001.

[27] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal Of Statistical Software*, 45(3):1–67, 2011.

[28] Zheng Wang, Ming-Jun Lai, Zhaosong Lu, Wei Fan, Hasan Davulcu, and Jieping Ye. Orthogonal Rank-One Matrix Pursuit for Low Rank Matrix Completion. *Proceedings of the 31st International Conference on Machine Learning*, pages 91–99., 2014.

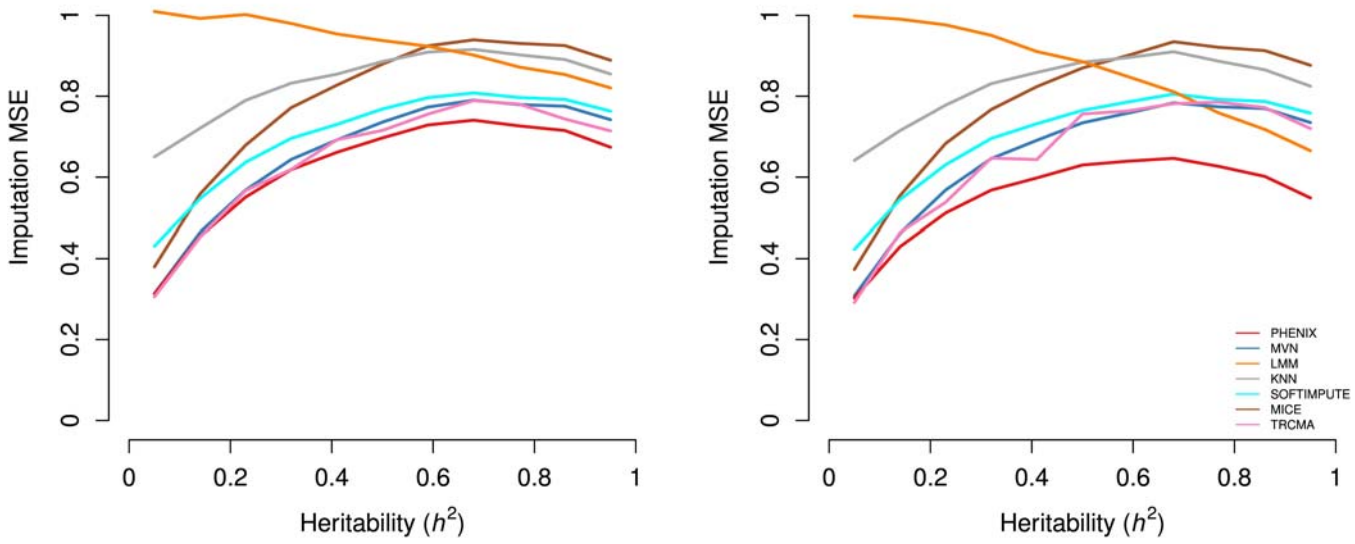
[29] BS Weir, AD Anderson, and AB Hepler. Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, 7(10):771–80, October 2006.

[30] Jianming Yu, Gael Pressoir, WH William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, Stephen Kresovich, Edward S Buckler, and IV Bi. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203–208, February 2006.

[31] Keyan Zhao, MJ Aranzana, Sung Kim, Clare Lister, Chikako Shindo, Chunlao Tang, Christopher Toomajian, Honggang Zheng, Caroline Dean, Paul Marjoram, and Magnus Nordborg. An Arabidopsis example of association mapping in structured samples. *PLoS genetics*, 3(1):e4, January 2007.

[32] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–4, July 2012.

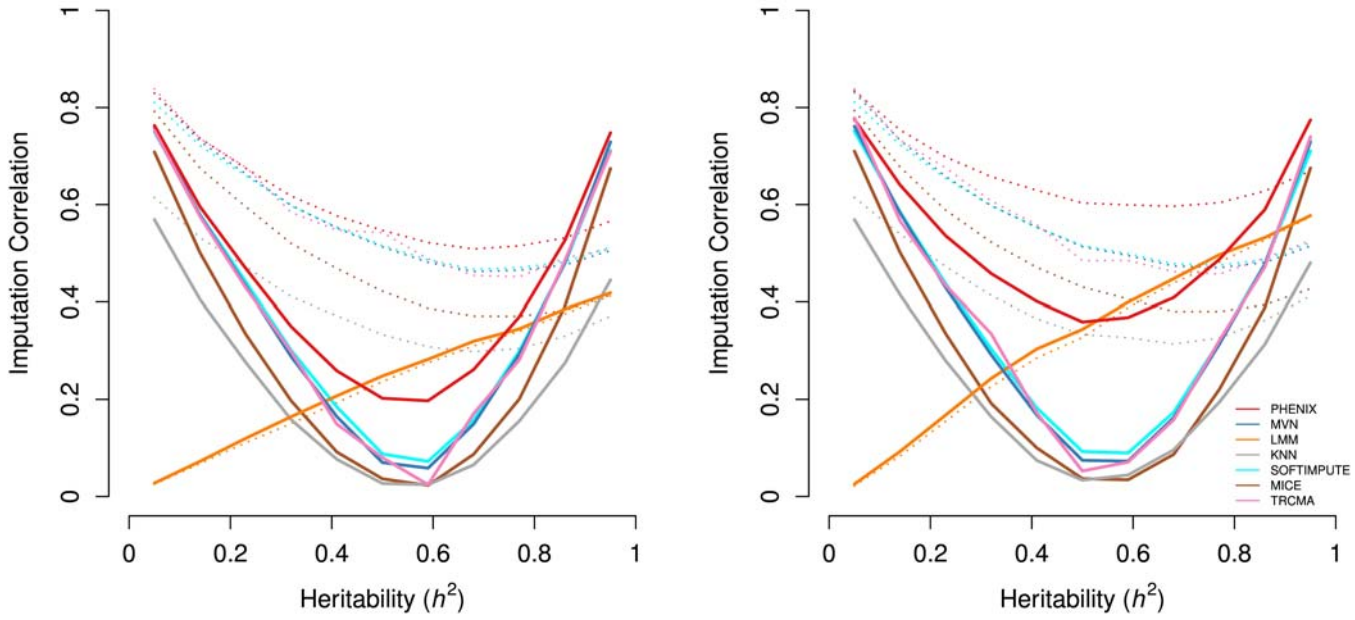
[33] Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4):407–409, February 2014.




Supplementary Figure 1

**Assessing phenotype imputation on simulated data using mean-squared error.**

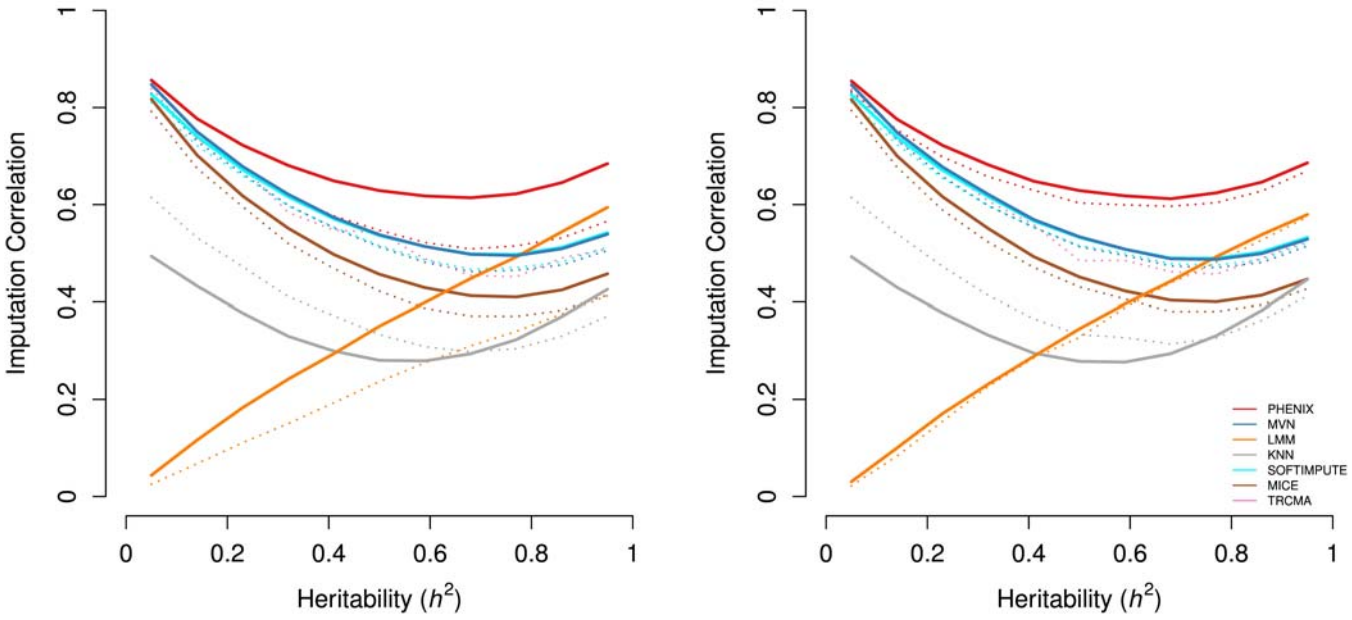
Simulation results measuring imputation accuracy with mean squared error (MSE) rather than correlation. Model 1: scenario simulated using an empirical kinship matrix derived from the human NSPHS study. Model 2: scenario simulated using 75 unrelated families of 4 sibs. Datasets were simulated at various levels of heritability (x-axis) for the traits. 300 individuals and 15 traits were simulated. 5% of phenotype values were set to missing before imputation. 7 different methods (legend) were applied to impute the missing values. The MSE between the imputed values and the true values is plotted on the y-axis for each method. Perfect imputation has MSE 0 and, because phenotypes are centered and standardized, imputing all entries to 0 has MSE 1. Compared to Figure 1, which uses correlation as an imputation metric, the results do not qualitatively change.




Supplementary Figure 2

**Cancellation of genetic and environmental covariances.**

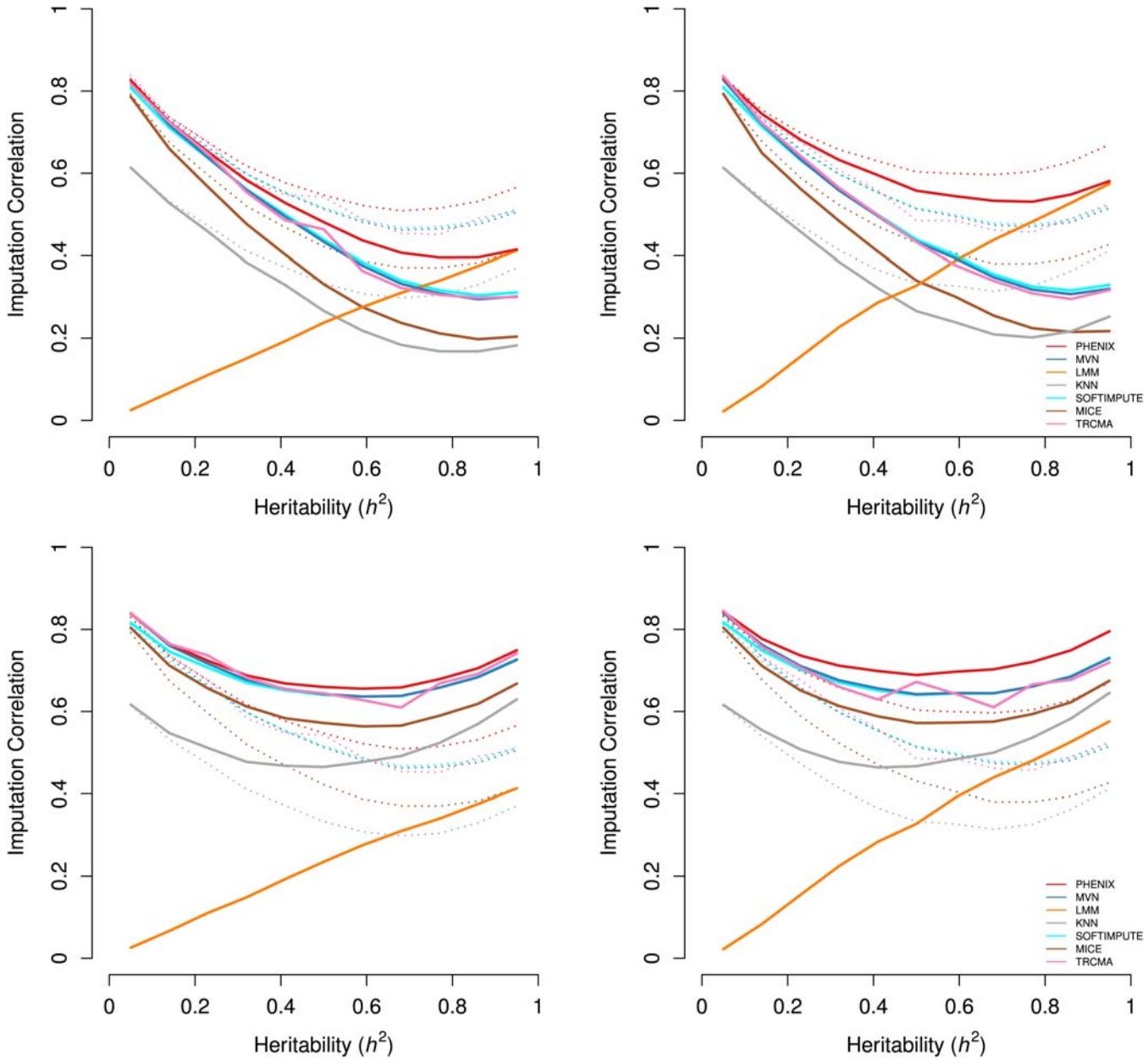
Simulation results with opposing genetic and environmental correlations. Rather than an AR matrix, this plot chooses genetic correlation $B$ to cancel the environmental correlation, $B_{pq} = -E_{pq}$ for $p \neq q$. 5% of phenotype values were held out and the correlation between the true and imputed values is plotted on the y-axis for each method. The dotted lines show the results from Figure 1 for reference.




Supplementary Figure 3

**Increasing sample size and number of phenotypes to *N*=1000, *P*=50.**

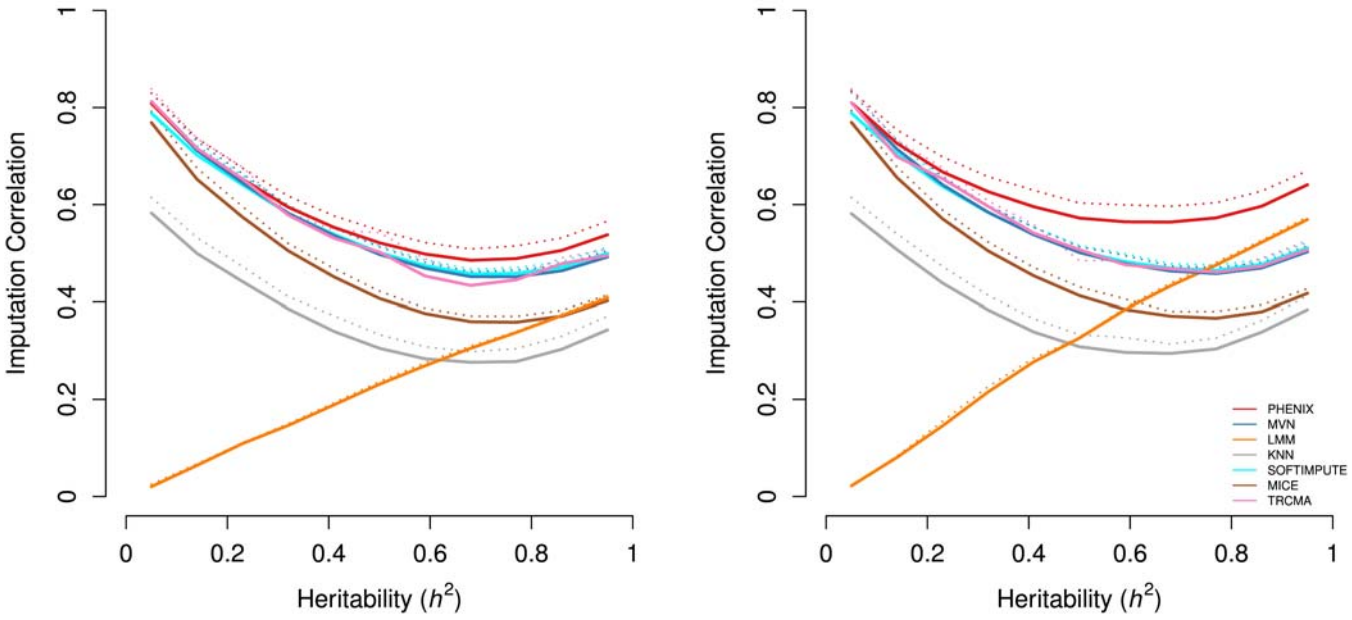
Simulation results using larger datasets. This figure uses (*N*,*P*)=(1000,50), while the dotted lines use (*N*,*P*)=(300,15). 5% of phenotype values were held out and the correlation between the true and imputed values is plotted on the y-axis for each method. Increasing the data size nearly always improves imputation accuracy, though this effect is attenuated when using the sibling relatedness matrix as family sizes are fixed and increasing *N* does not increase the amount of inter-sample correlation. The dotted lines show the results from Figure 1 for reference.




Supplementary Figure 4

**Varying levels of genetic correlation between phenotypes.**

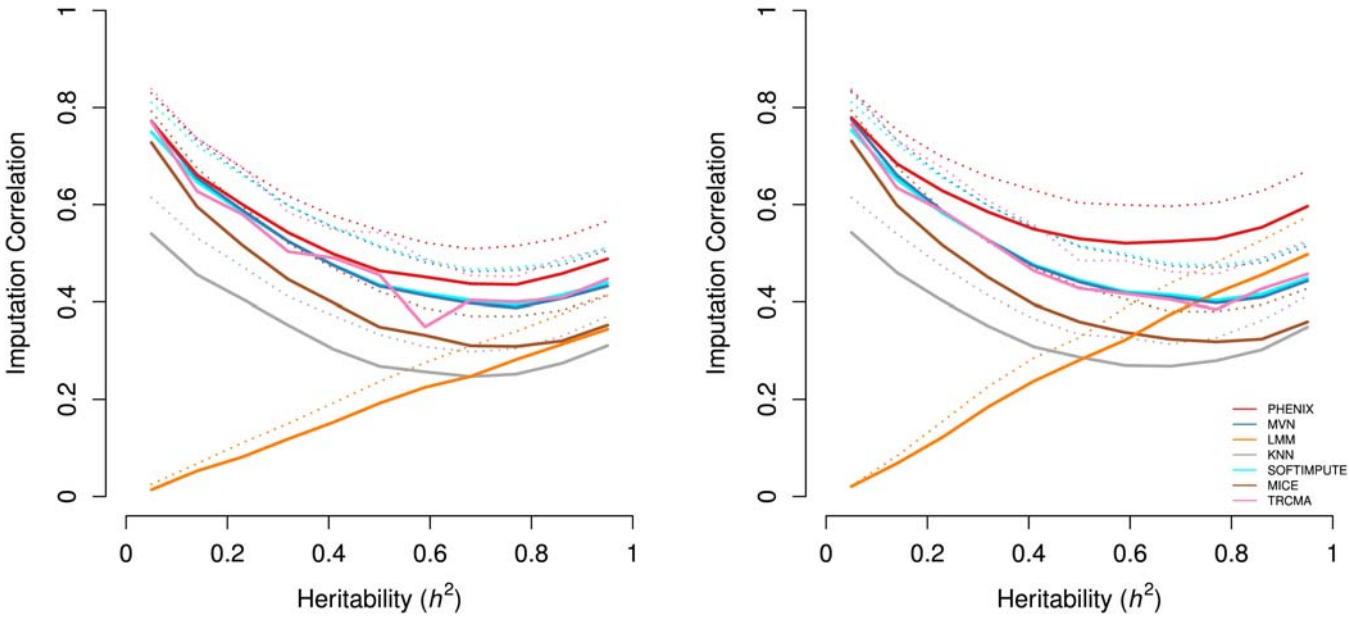
Simulation results varying the amount of genetic correlation. We vary the overall genetic correlation matrix *B* by changing ρ, the AR parameter. The top row shows simulations with ρ =.275, decreasing the average genetic correlation between traits compared to the dotted lines (from Figure 1) that use the baseline choice ρ =.45; the bottom row shows simulations with ρ = 0.675. Analogous results were obtained using ρ = -.275 (not shown). 5% of phenotype values were held out and the correlation between the true and imputed values is plotted on the y-axis for each method. Imputation accuracy of multitrait methods increases with genetic correlation and this effect increases with $h^2$.




Supplementary Figure 5

**Increasing data missingness to 10%.**

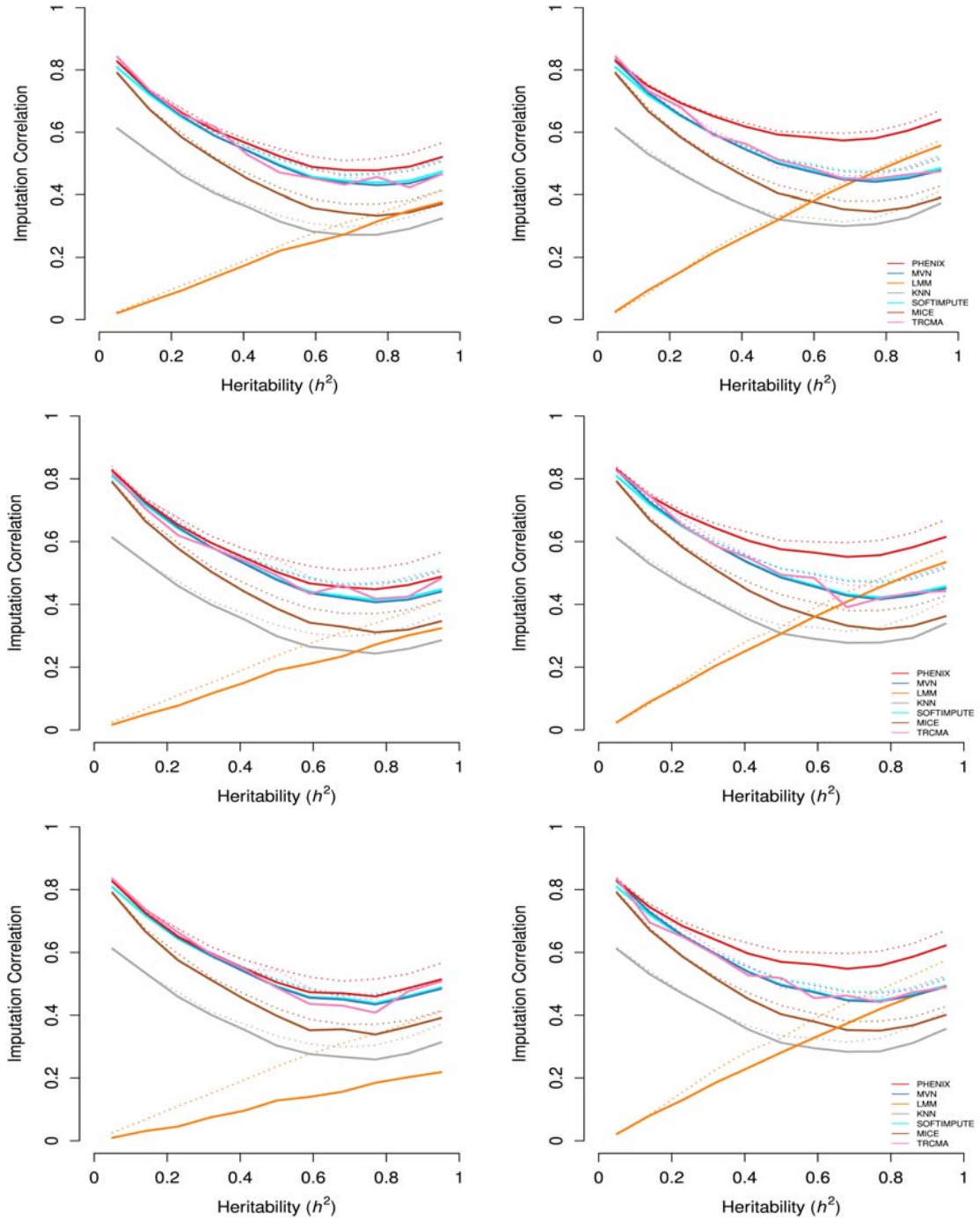
Simulation results at higher level of missingness. 10% of phenotype values were set to missing before imputation, rather than 5% as for the dotted lines. The correlation between the imputed values and the true values are plotted on the y-axis for each method. The dotted lines show the results from Figure 1 for reference.




Supplementary Figure 6

**Effect of non-random missingness.**

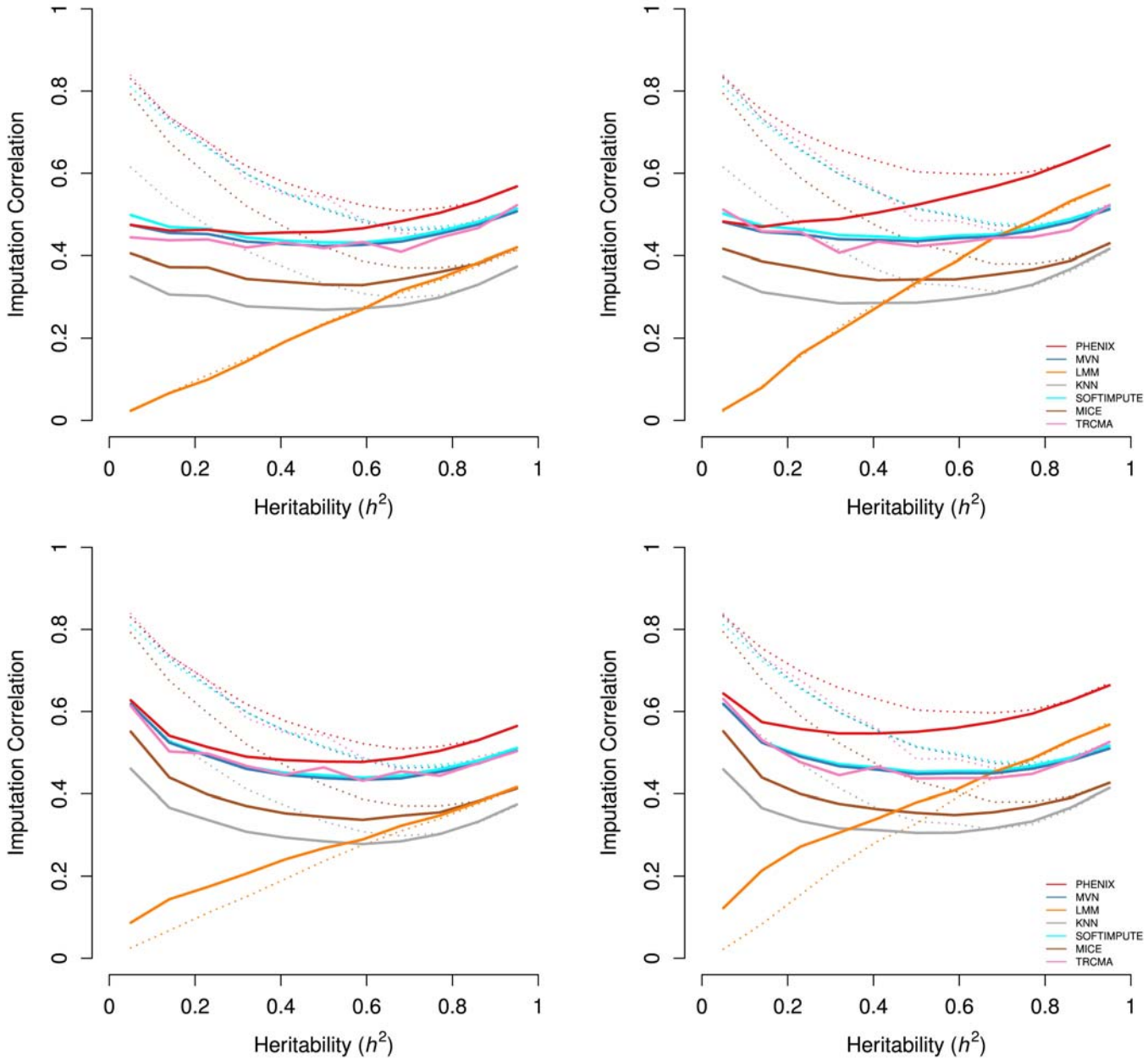
Simulation results with non-ignorable missingness. We hold out 5% of the entries of the phenotype matrix with probability increasing in their values and the correlation between the true and imputed values is plotted on the y-axis for each method. The dotted lines show the results from Figure 1 for reference.




Supplementary Figure 7

**Effect of unmodelled, shared environment.**

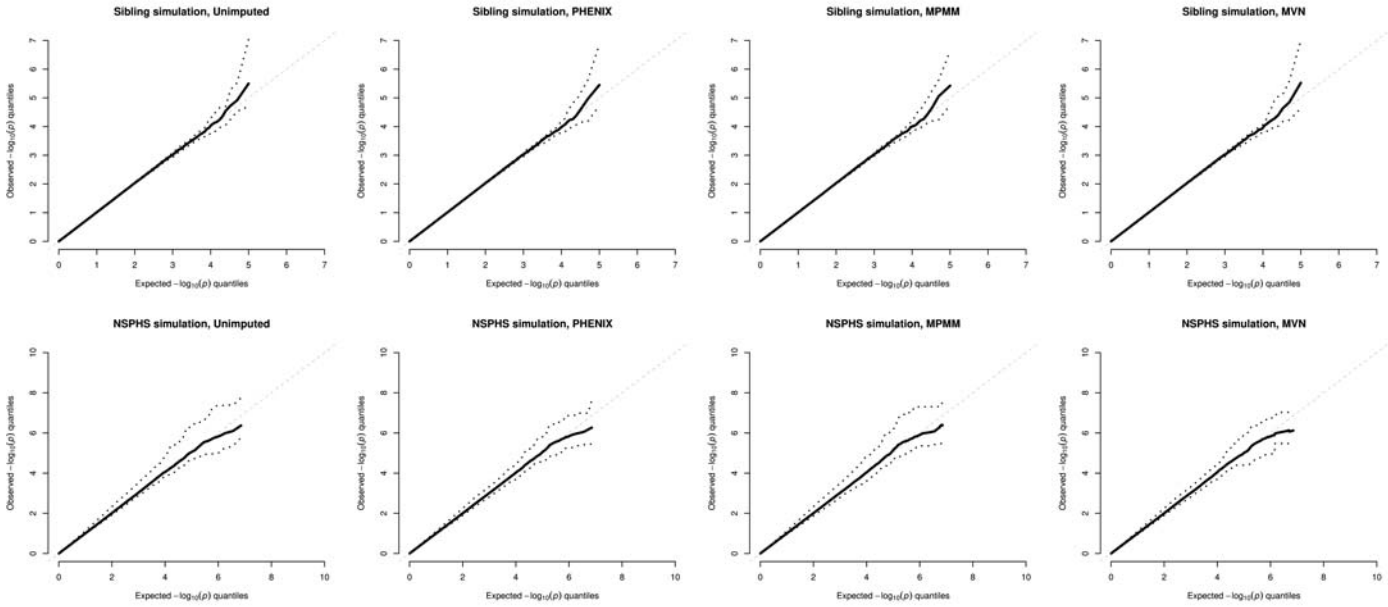
Simulation results with confounding cryptic relatedness. The contribution of the additive genetic term--*U*, in a typical MPMM--is $a^2$; each row increases the contribution of the contaminating shared environment, $c^2$, to the overall heritability, here defined as $h^2 = a^2 + c^2$. The first row uses $c^2 = .1a^2$; the second $c^2 = .3a^2$; and the last $c^2 = a^2$. 5% of phenotype values were held out and the correlation between the true and imputed values is plotted on the y-axis for each method. The dotted lines show the results from Figure 1 for reference.




Supplementary Figure 8

**Non-normally distributed phenotypes.**

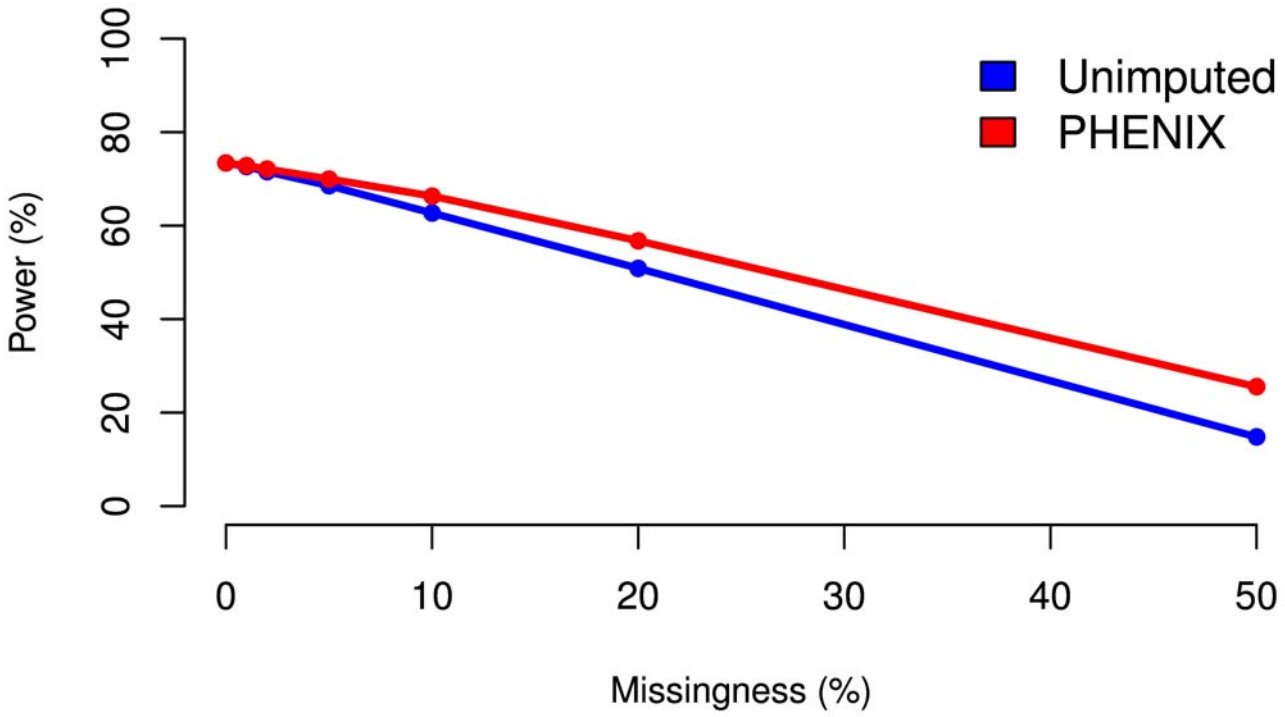
Simulation results with non-normal noise. We exponentially transform the environmental contribution, ε, to create log-normal noise. The resulting phenotypes are imputed without (top) or with (bottom) quantile normalization. 5% of phenotype values were held out and the correlation between the true and imputed values is plotted on the y-axis for each method. The dotted lines show the results from Figure 1 for reference.




Supplementary Figure 9

**Type I error calibration after phenotype imputation.**

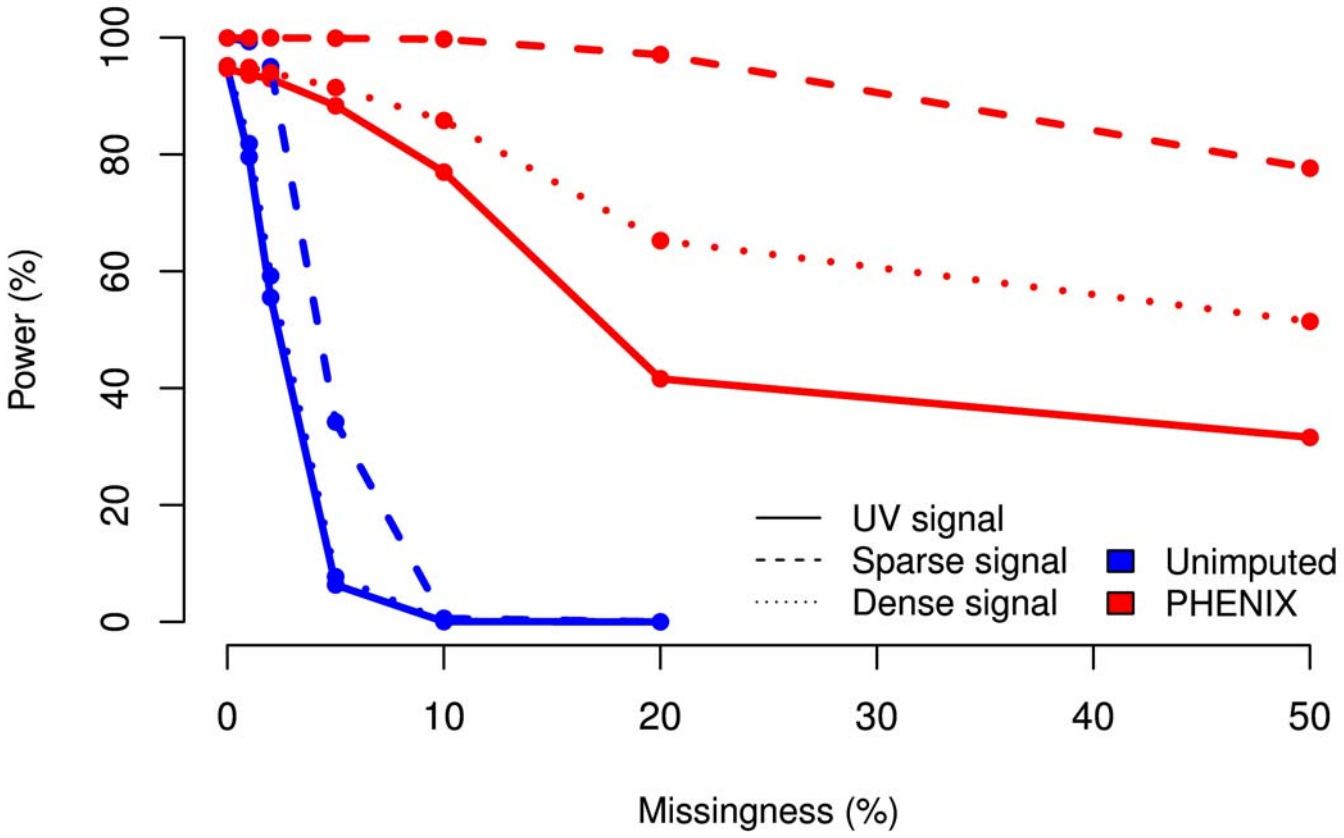
QQ plots from performing GWAS on 15 truly unassociated phenotypes with different imputation options (panel titles). Phenotypes are generated from our baseline simulation with the relevant *K* matrix. Rather than represent each of the 15 GWAS for each panel, we plot the point-wise minimum and maximum (dotted lines) and median (solid line) of the 15 lines. Top row: kinship and genotypes correspond to independent sets of 4 siblings. Bottom row: kinship and genotypes taken from NSPHS study.




Supplementary Figure 10

**Power of single phenotype tests after phenotype imputation.**

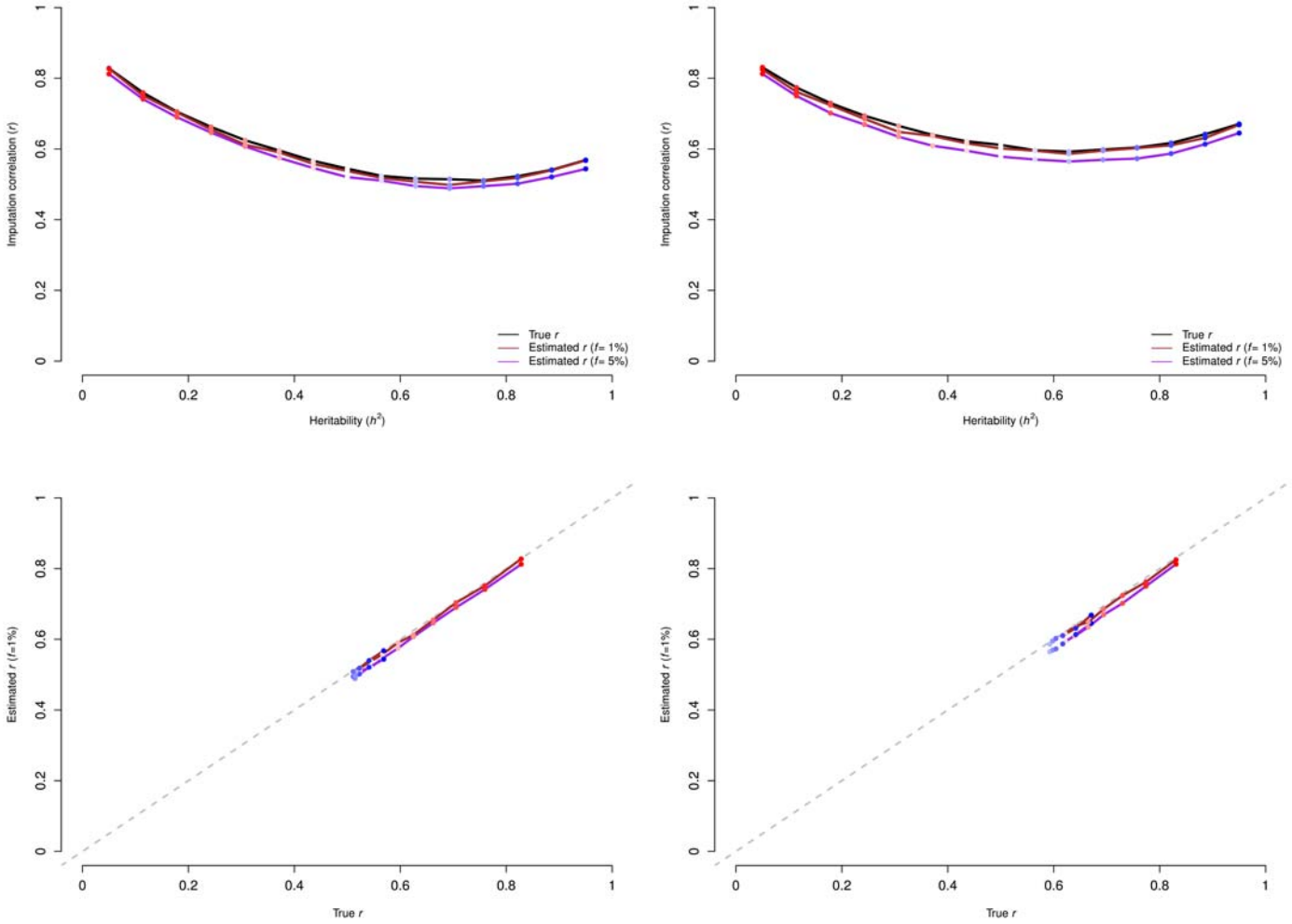
Power to detect a simulated, causal SNP using a univariate mixed model (LMM). 5,000 samples, comprising independent sets of 4 siblings, have 15 simulated phenotypes with pleiotropy. 5% of phenotypes are deleted and then an LMM is run with gemma after dropping missing data (Unimputed) or imputing with PHENIX. Power is calculated by averaging over 1,000 independently simulated datasets using the standard GWAS *p*-value threshold $5\times10^{-7}$.




Supplementary Figure 11

**Power of multiple phenotype tests after phenotype imputation.**

Power to detect a simulated, causal SNP using a multiphenotype mixed model (MPMM). 5,000 samples, comprising independent sets of 4 siblings, have 15 simulated phenotypes with three levels of pleiotropy (legend). 5% of phenotypes are deleted and then an MPMM is run with our method by dropping samples with any missing phenotype data (Unimputed) or imputing with PHENIX. Power is calculated by averaging over 5,000 independently simulated datasets using the standard GWAS *p*-value threshold $5\times10^{-7}$.




Supplementary Figure 12

**Calibration of the imputation metric *r.***

Calibration of our *r* metric for imputation accuracy. Data is from the baseline model, but we now record estimated imputation accuracies, which we call *r*, as well as the true imputation accuracies. Top row: imputation correlation is plotted against $h^2$. The black line is the true imputation accuracy and agrees with the PHENIX line (red) in Figure 1. We estimate *r* in two ways: by hiding 1% (brown line) or 5% (purple line) of observed entries. Point colors correspond to values of $h^2$. Bottom row: estimated *r* is compared to the true *r*, with variability created by varying $h^2$. Each point corresponds to the point in the below plot with the same color.